



CESAR

Central and Southeast European Resources

CIP-Pilot actions project no. 271022

www.cesar-project.net



Deliverable D3.3 (part A)
Third batch of resources:
documentation of the delivery

Version No. 1.0
2013-02-06

Document Information

Deliverable number:	D3.3 (part A)
Deliverable title:	Second batch of resources complying with the project's technical, linguistic, legal, etc. specifications
Due date of deliverable:	2013-01-31
Actual date of deliverable:	2013-02-06
Main author(s):	Maciej Ogrodniczuk (IPIPAN)
Participants:	<p>Tibor Pintér, Tamás Váradi, Dániel Varga, Veronika Vincze (HASRIL)</p> <p>András Balog, Mátyás Bartalis, Géza Németh, Gábor Olaszy, György Szaszák, Klára Vicsi (BME)</p> <p>Željko Agić, Božo Bekavac, Nikola Ljubešić, Sanda Martinčić-Ipšić, Ida Raffaelli, add Jan Šnajder, Krešo Šojat, Vanja Štefanec, Marko Tadić (FFZG)</p> <p>Łukasz Degórski, Katarzyna Głowińska, Anna Kibort, Łukasz Kobylński, Mateusz Kopeć, Katarzyna Krasnowska, Michał Lenart, Leszek Manicki, Małgorzata Marciniak, Marek Maziarz, Marcin Miłkowski, Bartłomiej Nitoń, Agnieszka Patejuk, Maciej Piasecki, Aleksander Pohl, Adam Przepiórkowski, Piotr Przybyła, Agata Savary, Filip Skwarski, Jan Szejko, Joanna Świetlicka, Zygmunt Vetulani, Adam Wardyński, Jakub Waszczuk, Dawid Weiss, Marcin Woliński, Alina Wróblewska, Bartosz Zaborowski, Marcin Zajac (IPIPAN)</p> <p>Piotr Pęzik, Łukasz Drózd, Maciej Buczek (ULodz)</p> <p>Cvetana Krstev, Zoran Lučić, Miljana Milanović, Mirko Spasić, Ranka Stanković, Miloš Utvić, Duško Vitas, Anđelka Zečević (UBG)</p> <p>Mirko Spasić, Natalija Kovačević, Uroš Milošević, Jelena Jovanović, Nikola Dragičević, Mladen Stanojević (PUPIN)</p> <p>Tsvetana Dimitrova, Svetla Koeva (IBL)</p> <p>Radovan Garabík, Adriána Žáková (LSIL)</p>
Workpackage:	WP3
Workpackage title:	Enhancing language resources
Workpackage leader:	IPIPAN

Dissemination level:	PU: Public
Version:	1.0
Keywords:	language resources and technologies, LR, LRT, upgraded resources, extended resources, cross-lingually linked resources

History of Versions

Ver- sion	Date	Author (Partner)	Contributions	Description/Approval Level
1.0	2013-02-06	Maciej Ogrodniczuk	All partners (listed above as participants)	Automatically generated content basing on metadata descriptions contributed by partners.

EXECUTIVE SUMMARY

This deliverable provides a complete documentation of the delivery of resources uploaded by CESAR consortium partners to META-SHARE CESAR node throughout the whole duration of the project (until end of January 2013).

The resources have been documented according to XML Schema model provided by META-NET and fully reflect information available in the META-SHARE metadata.

Table of Contents

1 HASRIL resources.....	11
1.1 Szeged Corpus	11
1.2 Szeged Treebank	12
1.3 Szeged Named Entity Recognition Corpus	14
1.4 Hungarian WordNet	16
1.5 Hungarian Webcorpus	17
1.6 Hunglish Corpus.....	18
1.7 morphdb.hu.....	20
1.8 hunmorph.....	21
1.9 hunalign	22
1.10 huntoken	24
1.11 Hungarian Opinion-Tagged Sentence Bank	25
1.12 HunNERwiki: Automatically generated NE tagged corpus for Hungarian.....	32
1.13 Hungarian Verb Phrase Constructions	35
1.14 hunner.....	37
1.15 hunpars	39
1.16 hunpos	41
1.17 Hungarian WSD Corpus.....	43
1.18 Hungarian Language Processing Tools in NooJ.....	46
1.19 SzegedParalell	49
1.20 Szeged Criminal NE Corpus	51
1.21 SzegedParalellFX	53
1.22 Szeged Treebank FX	56
1.23 Hungarian National Corpus.....	58
1.24 HuComTech Multimodal Corpus and Database.....	60
1.25 BEA Hungarian spontaneous speech database	62
1.26 Hungarian Kindergarten Language Corpus	65
1.27 ht-online.....	68
1.28 Hungarian Consize Dictionary	70
1.29 HHC: Hungarian historical corpus	72
1.30 N-grams from Hungarian National Corpus	75
1.31 CHSM-IC: Corpus of Hungarian School Metalanguage – Interview Corpus	77
2 BME-TMIT resources.....	79
2.1 Mindentudás Speech Corpus	79
2.2 Word level speech database for Hungarian	81
2.3 Hungarian BABEL	87
2.4 Hungarian Broadcast News Database.....	92
2.5 Sound Gesture Database.....	96
2.6 Hungarian Speech Emotion Database	99
2.7 Hungarian MTBA.....	104
2.8 Hungarian MRBA	108
2.9 Hungarian Phone Speech Call Center Database.....	112

2.10 Hungarian BABEL phonetic segmentation and syntactic and prosodic analysis	116
2.11 Di-phone database for text-to-speech conversion.....	119
2.12 Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection	125
2.13 Read speech database in Hungarian	131
2.14 Hungarian Book (Egri csillagok/Eclipse of the Crescent Moon by Géza Gárdonyi) Reading Speech and Aligned	138
2.15 Hungarian Poem (János vitéz/John the Valiant by Sándor Petőfi) Reading Speech and Aligned	141
2.16 Hungarian Parliamentary Speech and Aligned Text Selection Database	144
2.17 Named entity lexical database	147
2.18 Formant database from spoken words	152
2.19 Graphical query interface for Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection	158
2.20 Hungarian Medical Speech Database	162
2.21 Automatic Prosodic Segmenter	166
2.22 Hungarian Phonetic Transcriber.....	169
2.23 Hungarian MALACH Database	171
2.24 Hungarian Broadcast Conversation Database from the Catholic Radio of Eger (EKR)	174
2.25 Accent marker database for Hungarian written sentences.....	176
3 FFZG resources.....	179
3.1 Croatian National Corpus v2.5	180
3.2 Croatian Morphological Lexicon v4.6.....	183
3.3 Croatian-English Parallel Corpus	186
3.4 Croatian Lemmatisation Server	189
3.5 Croatian Valency Lexicon.....	192
3.6 Croatian Web Corpus	196
3.7 Slovene Web Corpus	198
3.8 Croatian-English Parallel Web Corpus.....	201
3.9 South-East European Parallel Corpus.....	204
3.10 Croatian Dependency Treebank	208
3.11 Web Content Extractor	211
3.12 Collocation and Term Extractor	214
3.13 Croatian Language Web Services.....	217
3.14 Croatian Translations of Acquis.....	220
3.15 Croatian National Corpus v3.0	223
3.16 Corpus of Narodne novine.....	226
3.17 Croatian n-grams	229
3.18 Croatian Morphological Lexicon v5.0.....	232
3.19 Orwell 1984 Croatian	235
3.20 Croatian WordNet	238
3.21 Croatian Automatic Collocations Dictionary	242
3.22 Croatian Weather Dialogue Corpus.....	244
3.23 CESAR Aligned Wikipedia Headwords List	248
3.24 Croatian and Slovene NERC models for Stanford NERC.....	251

3.25 Coral Corpus Aligner	255
3.26 Croatian Sentiment Lexicon	258
4 IPIAN resources	261
4.1 Polish Sejm Corpus	261
4.2 PoliMorf Inflectional Dictionary	268
4.3 Polish WordNet	272
4.4 Polish Named Entity Recognition Tool.....	276
4.5 1 million subcorpus of National Corpus of Polish	281
4.6 Polish Named Entity Gazetteer.....	287
4.7 LUNA.PL Corpus.....	293
4.8 LUNA-WOZ.PL Corpus	295
4.9 Morphosyntactic tagset converter for positional tagsets	297
4.10 Spejd.....	301
4.11 N-grams from balanced National Corpus of Polish.....	305
4.12 Distributable subcorpus of National Corpus of Polish	307
4.13 Morfeusz Polimorf.....	312
4.14 Morfologik Inflectional Dictionary	317
4.15 Grammatical Lexicon of Polish Phraseology	321
4.16 Grammatical Lexicon of Polish Economical Phraseology	325
4.17 Grammatical Lexicon of Warsaw Urban Proper Names	329
4.18 Multilingual lexicon of toponyms	333
4.19 Polish Valence Dictionary.....	338
4.20 Summarizer.....	340
4.21 morfologik-stemming	342
4.22 Corpus of the Polish language of the 1960s	345
4.23 Shallow Grammar for the National Corpus of Polish.....	350
4.24 PANTERA.....	354
4.25 PolNet – Polish Wordnet v.1	359
4.26 Polish Wikipedia Corpus.....	363
4.27 SEJFEK4Spejd	365
4.28 PNET	369
4.29 An LFG grammar of Polish (POLFIE).....	372
4.30 Lexeme Forge.....	375
4.31 A tool for creating a Polish Valence Dictionary.....	377
4.32 CorpCor	380
4.33 plWikiEcono.....	386
4.34 plWikiEconoSenses	388
4.35 Prolexbase	390
4.36 Dependency Parsing Model for Polish	396
4.37 Polish HateSpeech Corpus	399
4.38 Polish Coreference Corpus	401
4.39 Polish Coreference Tools	407
4.40 Syntactic-Generative Dictionary of Polish Verbs.....	409
4.41 Manually aligned CES Polish-English parallel corpus.....	413

4.42 Polish lexicon for OpenCyc.....	419
4.43 TAG grammar for Polish.....	422
4.44 Anotatornia.....	424
4.45 Constrained Conditional Random Fields Tagging Tool.....	427
4.46 Multiservice.....	431
4.47 DBpedia resource classification into the OpenCyc taxonomy	434
4.48 DBpedia Extender	436
4.49 LFG Treebank for Polish.....	439
4.50 NKJP1mEcono corpus.....	441
4.51 gpwEcono corpus	445
4.52 Summary Annotation Tools	449
4.53 DistSys.....	451
4.54 The Polish SRL corpus.....	452
4.55 Składnica — a treebank of Polish.....	453
4.56 Świgra — a DCG parser of Polish.....	456
4.57 NKJP Model for TnT Tagger for Polish.....	458
4.58 Polish Automatic Collocations Dictionary	461
4.59 Polish Corpus of Wrocław University of Technology	462
5 ULodz resources	464
5.1 PELCRA Polish-English parallel corpora (CC-BY)	464
5.2 PELCRA Polish-English parallel corpora (CC-BY-NC)	475
5.3 PELCRA Polish spoken corpus (CC-BY-NC)	481
5.4 ECL Dictionaries	487
5.5 PELCRA EN Lemmatizer	492
5.6 PELCRA Language Detector	497
5.7 PELCRA Polish-English parallel corpus of literary works (CC-BY)	501
5.8 PELCRA multilingual parallel corpora (CC-BY)	508
5.9 OSW Polish-English parallel corpus (CC-BY-NC)	532
5.10 PELCRA time-aligned spoken corpus of Polish (CC-BY-NC).....	539
5.11 PELCRA WebLign crawler.....	548
5.12 PELCRA Word Aligned Corpora.....	553
5.13 Spelling and NUmbers Voice database	557
5.14 HASK collocation dictionary (English)	564
5.15 HASK collocation dictionary (Polish).....	569
5.16 PELCRA Spoken Learner English Corpus.....	573
5.17 Polish-Russian Parallel Corpus	582
6 UBG resources	588
6.1 Serbian Wordnet.....	588
6.2 Corpus of Contemporary Serbian	593
6.3 Serbian Lemmatized and PoS Annotated Corpus.....	597
6.4 French-Serbian Aligned Corpus	602
6.5 Multilingual Edition of Verne's Novel "Around the World in 80 Days".....	606
6.6 Organizing digitized material	612

6.7 English-Serbian Aligned Corpus	615
6.8 Serbian NooJ module	620
6.9 Serbian Morphological Dictionary (Multext-East).....	625
6.10 Morphosyntactically tagged Serbian version of Jules Verne's novel "Around the world in 80 days"	628
6.11 Bibliša: Aligned Collection Search Tool.....	633
6.12 Corpus of Contemporary Serbian Newspapers and Magazines	636
6.13 Named Entities evaluation corpus for Serbian	640
6.14 Serbian NGrams	644
6.15 NERanka: Named Entity Recognition and Annotation Tool	647
6.16 Serbian Spell Checker	651
6.17 Emotions Annotation Tool	655
6.18 NERosetta.....	658
6.19 Rhetorical Figures for Serbian.....	661
7 PUPIN resources	664
7.1 MONO version of NooJ	664
7.2 Java version of NooJ	667
8 IBL resources	671
8.1 Bulgarian National Corpus	671
8.2 The Bulgarian National Corpus Collocation service	674
8.3 Bulgarian Part-of-Speech Corpus	675
8.4 Bulgarian Sense-Annotated Corpus.....	678
8.5 Bulgarian-X language Parallel Corpus	681
8.6 Bulgarian WordNet	689
8.7 Bulgarian WordNet - web access	694
8.8 Bulgarian Spell Checker for Windows.....	697
8.9 Bulgarian Spell Checker Web Service	700
8.10 The Bulgarian-X Language Parallel Corpus Collocations service	702
8.11 Lists of Bulgarian Multiword Expressions	703
8.12 Bulgarian Frequency Dictionaries	705
8.13 Hydra - tool for developing wordnets.....	708
8.14 Chooser - annotation tool	711
8.15 Bulgarian Sentence Splitter and Tokenizer	714
8.16 Web based infrastructure for Bulgarian data processing	716
8.17 TREFL – Translation Reference Library.....	720
8.18 SARP - Speech Analyzer Rapid Plot. Plotting vowels in F2-F1 scatter charts with multiple data sets.....	723
8.19 RTComp - Real Time Comparison.....	725
8.20 Corpus of Spoken Bulgarian	727
8.21 Corpus of Colloquial Bulgarian.....	730
8.22 Dictionary of Synonyms in Bulgarian Language	733
8.23 Dictionary of Antonyms in Bulgarian Language	736
8.24 Register of Phraseologisms in Bulgarian Language.....	739
8.25 Dictionary of Neologisms in Bulgarian Language	742

8.26 Bulgarian Spell Checker for Mac OS	745
8.27 Wiki1000+ corpus with annotated MWEs	747
8.28 The Bulgarian-English Sentence- and Clause-Aligned Corpus.....	749
8.29 Multilingual dictionaries.....	752
8.30 Bulgarian MWE dictionary	755
8.31 bgMWE – tool for MWE recognition.....	757
8.32 TextMatch	759
8.33 Bulgarian Grammar checker web service.....	761
8.34 N-grams from Bulgarian National Corpus	764
8.35 Bulgarian Automatic Collocations Dictionary	766
8.36 Bibliography of Bulgarian Lexicology, Phraseology and Lexicography	768
9 LSIL resources	773
9.1 Slovak National Corpus prim-5.0.....	773
9.2 Corpus of Spoken Slovak	775
9.3 Slovak Morphology Database	779
9.4 Slovak-Czech Parallel Corpus (all)	780
9.5 Slovak-English Parallel Corpus (all)	783
9.6 Slovak Treebank.....	785
9.7 Balanced Slovak Corpus prim-5.0-vyv	786
9.8 Manually Annotated Slovak Corpus.....	788
9.9 Language model prim-5.0-sane	791
9.10 Language model prim-5.0-inf.....	792
9.11 Language model prim-5.0-vyv	794
9.12 Corpus of Legal Texts	796
9.13 Slovak Web Corpus	798
9.14 Slovak-Czech Parallel Corpus (free)	800
9.15 Slovak-English Parallel Corpus (free).....	802
9.16 Slovak Terminology Database.....	804
9.17 Corpus of Informational Texts prim-6.0-inf.....	806
9.18 Corpus of Professional Texts prim-6.0-prf.....	808
9.19 Corpus of Fiction prim-6.0-img	810
9.20 Corpus of Original Slovak Texts prim-6.0-sk	812
9.21 Corpus of Original Slovak Fiction skimg-6.0	814
9.22 Corpus of Slovak Texts from the Years 1955 to 1989 R55AZ89	816
9.23 Corpus of Historical Slovak	818
9.24 Lietuvių kalbos WordNet (Lithuanian WordNet).....	820
9.25 Slovník slovných spojení v slovenčine. Podstatné mená (Dictionary of Slovak Collocations. Nouns).....	822
9.26 Slovník slovných spojení v slovenčine. Prídavné mená (Dictionary of Slovak Collocations. Adjectives)	824
9.27 Slovak WordNet.....	825
9.28 Slovak National Corpus prim-6.0.....	827
9.29 Balanced Slovak Corpus prim-6.0-vyv	829
9.30 Language model prim-6.0-sane	831

9.31 Language model prim-6.0-inf.....	833
9.32 Language model prim-6.0-vyv	835
9.33 Parallelum Slovaco-Latinum Corpus.....	836
9.34 n-grams from Slovak National Corpus.....	838
9.35 Multilingual Glossary of Synsets	839
9.36 Automatic Collocation Dictionary of Slovak	841

Introduction

This deliverable provides a documentation of the delivery of resources made available in META-SHARE by CESAR consortium partners until the end of the project (January 2013). The resources have been documented according to XML Schema model provided by META-NET and fully reflect information available in the META-SHARE metadata. This document was generated automatically from resource descriptions provided by CESAR partners.

1. HASRIL resources

1.1. Szeged Corpus

General Information

Description	A morpho-syntactically annotated and manually disambiguated corpus of 1.2 million words.
Identifier	101
Resource type	Corpus
URL	http://www.inf.u-szeged.hu/rgai/SzegedTreebank
Version	2.0

Contacts

János Csirik	
Contact	csirik@inf.u-szeged.hu
Organization	University of Szeged Department of Informatics csirik@inf.u-szeged.hu
Veronika Vincze	
Contact	vinczev@inf.u-szeged.hu
Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu
Richárd Farkas	
Contact	rfarkas@inf.u-szeged.hu
Organization	University of Szeged Department of Informatics rfarkas@inf.u-szeged.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

MS-NC-NoReD		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Signatories	Veronika Vincze	
	Position	research fellow
	Contact	Tisza Lajos krt. 103. 6720 Szeged vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html
	Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu

Metadata

Creation date	2011-10-17	
Metadata creators	Veronika Vincze	
	Contact	vinczec@inf.u-szeged.hu
	Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Size	82000 sentences	
Annotation	Morphosyntactic annotation – POS tagging	
	Segmentation level	Word

1.2. Szeged Treebank

General Information

Description	A manually checked treebank of 1.2 million words.
Identifier	102
Resource type	Corpus
URL	http://www.inf.u-szeged.hu/rgai/nlp/SzegedTreebank

Version	2.0
----------------	-----

Contacts

János Csirik	
Contact	csirik@inf.u-szeged.hu
Organization	University of Szeged Department of Informatics csirik@inf.u-szeged.hu
Veronika Vincze	
Contact	vinczev@inf.u-szeged.hu
Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu
Richárd Farkas	
Contact	rfarkas@inf.u-szeged.hu
Organization	University of Szeged Department of Informatics rfarkas@inf.u-szeged.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

MS-NC-NoReD		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Signatories	Veronika Vincze	
	Position	research fellow
	Contact	Tisza Lajos krt. 103. 6720 Szeged vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html
	Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu

Metadata

Creation date	2011-10-17

Metadata creators	Veronika Vincze	
	Contact	vinczev@inf.u-szeged.hu
	Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Size	82000 sentences	
Annotation	Morphosyntactic annotation – POS tagging	
	Segmentation level	Word
	Syntactic annotation – treebanks	
	Segmentation level	Word group

1.3. Szeged Named Entity Recognition Corpus

General Information

Short name	Szeged NER Corpus
Description	The Szeged NER corpus is a manually annotated part of the Szeged Treebank, consisting of short business news. The used NER categories are (based on the CoNLL system (http://www.cnts.ua.ac.be/conll2003/ner/)) the following: PERSON, ORGANISATION, LOCATION and OTHER.
Identifier	103
Resource type	Corpus
URL	http://www.inf.u-szeged.hu/rgai/corpus_ne
Version	1.0

Contacts

János Csirik	
Contact	csirik@inf.u-szeged.hu
Organization	University of Szeged Department of Informatics csirik@inf.u-szeged.hu
Veronika Vincze	

Contact	vinczev@inf.u-szeged.hu
Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu
Richárd Farkas	
Contact	rfarkas@inf.u-szeged.hu
Organization	University of Szeged Department of Informatics rfarkas@inf.u-szeged.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

MS-NC-NoReD		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://www.inf.u-szeged.hu/rgai/corpus_ne	
Signatories	Veronika Vincze	
	Contact	vinczev@inf.u-szeged.hu
	Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu

Metadata

Creation date	2011-10-17	
Metadata creators	Veronika Vincze	
	Contact	vinczev@inf.u-szeged.hu
	Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Size	200000 tokens	

Annotation	Semantic annotation – named entities	
	Segmentation level	Word group

1.4. Hungarian WordNet

General Information

Short name	HuWN
Description	The Hungarian WordNet is a multilingual ontology, meaning that most of its synsets were mapped to equivalent concepts in English (Princeton) WordNet v. 2.0. The ontology is also linked to entries of a Hungarian Monolingual explanatory dictionary and to the entries of the Hungarian verb valency frame lexicon.
Identifier	104
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Wordnet
URL	http://www.inf.u-szeged.hu/rgai/HuWN
Version	1.0

Contacts

János Csirik	
Contact	csirik@inf.u-szeged.hu
Organization	University of Szeged Department of Informatics csirik@inf.u-szeged.hu
Veronika Vincze	
Contact	vinczev@inf.u-szeged.hu
Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu
Richárd Farkas	
Contact	rfarkas@inf.u-szeged.hu
Organization	University of Szeged Department of Informatics rfarkas@inf.u-szeged.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

MS-NC-NoReD	
Restrictions of use	Academic – non-commercial use
Access medium	Downloadable

Metadata

Creation date	2011-10-17	
Metadata creators	Veronika Vincze	
	Contact	vinczev@inf.u-szeged.hu
	Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu

Lexical conceptual resource

Lexical conceptual resource type	Wordnet
---	---------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Size	42000 synsets	

1.5. Hungarian Webcorpus

General Information

Description	A Hungarian gigacorpora scraped from the .hu domain.
Identifier	105
Resource type	Corpus

Contacts

Dániel Varga	
Contact	daniel@mokk.bme.hu

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

CC-BY	
Restrictions of use	Other
Access medium	Downloadable
Download location	http://mokk.bme.hu/resources/webcorpus

Metadata

Creation date	2011-11-29	
Metadata creators	Dániel Varga	
	Contact	daniel@mokk.bme.hu
	Organization	Budapest University of Technology and Economics daniel@mokk.bme.hu

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Size	589000000 tokens	
Annotation	Other	
	Segmentation level	Sentence

1.6. Hunglish Corpus

General Information

Description	Hungarian-English parallel corpus automatically aligned at the sentence level.
Identifier	106
Resource type	Corpus
Version	2.0

Contacts

Dániel Varga	
Contact	daniel@mokk.bme.hu
Organization	Budapest University of Technology and Economics MOKK Centre for Media Research and Education

	daniel@mokk.bme.hu
--	--

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

CC-BY		
Restrictions of use	Other	
Access medium	Downloadable	
Download location	http://mokk.bme.hu/resources/hunglishcorpus	
Signatories	Dániel Varga	
	Contact	daniel@mokk.bme.hu
	Organization	Budapest University of Technology and Economics MOKK Centre for Media Research and Education daniel@mokk.bme.hu

Metadata

Creation date	2011-11-12	
Metadata creators	Dániel Varga	
	Contact	daniel@mokk.bme.hu
	Organization	Budapest University of Technology and Economics MOKK Centre for Media Research and Education daniel@mokk.bme.hu

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Comparable	
Multilinguality type details	Automatically sentence-segmented and aligned.	
Languages	Hungarian	
	Language ID	HU
	English	
	Language ID	EN
Size	2000000 sentences	
Annotation	Segmentation	
	Segmentation	Sentence

	level	
--	-------	--

1.7. morphdb.hu

General Information

Description	Hungarian lexical database and morphological grammar
Identifier	107
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Lexicon

Contacts

Dániel Varga	
Contact	daniel@mokk.bme.hu

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

CC-BY	
Restrictions of use	Other
Access medium	Downloadable
Download location	http://mokk.bme.hu/resources/morphdb-hu

Metadata

Creation date	2011-11-13	
Metadata creators	Dániel Varga	
	Contact	daniel@mokk.bme.hu
	Organization	Budapest University of Technology and Economics daniel@mokk.bme.hu

Lexical conceptual resource

Lexical conceptual resource type	Lexicon
---	---------

Texts

Media type	text
-------------------	------

Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Size	400000 items	

1.8. hunmorph

General Information

Short name	hunmorph
Description	hunmorph is an open source tool and programming library for stemming and morphological analysis.
Identifier	108
Resource type	Tool/service
Tool/service type	Tool
URL	http://mokk.bme.hu/resources/hunmorph
Version	1.0
Last update	2011-10-11

Contacts

Varga Daniel	
Contact	daniel@mokk.bme.hu
Organization	Budapest University of Technology and Economics Media Research Centre daniel@mokk.bme.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

LGPL		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://mokk.bme.hu/resources/hunmorph	
Fee	free of charge	
Distribution rights holder	Budapest University of Technology and Economics	
	Short name	BUTE MOKK

	Department name	Media Research Centre
	Contact	daniel@mokk.bme.hu

Metadata

Creation date	2011-11-30
----------------------	------------

Usage

Foreseen use	NLP applications
NLP-specific use	Morphological analysis

Resource creation

Creation start date	2010-03-01
----------------------------	------------

Tool/service

Tool/service type	Tool	
Language dependent	False	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	Linux	
Required software	OcaML	
Tool/service creation	Implementation language	OcaML
	Formalism	suffix stripping

1.9. hunalign

General Information

Short name	hunalign
Description	hunalign is a sentence aligner. It can use bilingual lexicons as a resource, but in the lack of such lexicon, its automatic lexicon-builder ensures that its precision degrades only marginally.
Identifier	109
Resource type	Tool/service

Tool/service type	Tool
URL	http://mokk.bme.hu/resources/hunalign
Version	1.0
Last update	2011-10-11

Contacts

Varga Daniel	
Contact	daniel@mokk.bme.hu
Organization	Budapest University of Technology and Economics Media Research Centre daniel@mokk.bme.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

LGPL		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://mokk.bme.hu/resources/hunalign	
Fee	free of charge	
Distribution rights holder	Budapest University of Technology and Economics	
	Short name	BUTE MOKK
	Department name	Media Research Centre
	Contact	daniel@mokk.bme.hu

Metadata

Creation date	2011-11-30
----------------------	------------

Usage

Foreseen use	NLP applications
NLP-specific use	Bilingual lexicon induction

Resource creation

Creation start date	2004-08-01
----------------------------	------------

Tool/service

Tool/service type	Tool	
Language dependent	False	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	Linux Windows	
Tool/service creation	Implementation language	C++

1.10. huntoken

General Information

Short name	huntoken
Description	huntoken is an open source tool for tokenization and sentence segmentation.
Identifier	110
Resource type	Tool/service
Tool/service type	Tool
URL	http://mokk.bme.hu/resources/huntoken
Version	1.0
Last update	2005-10-11

Contacts

Varga Daniel	
Contact	daniel@mokk.bme.hu
Organization	Budapest University of Technology and Economics Media Research Centre daniel@mokk.bme.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

LGPL

Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://mokk.bme.hu/resources/huntoken	
Fee	free of charge	
Distribution rights holder	Budapest University of Technology and Economics	
	Short name	BUTE MOKK
	Department name	Media Research Centre
	Contact	daniel@mokk.bme.hu

Metadata

Creation date	2011-11-30
----------------------	------------

Usage

Foreseen use	NLP applications
NLP-specific use	Other

Resource creation

Creation start date	2003-09-01
----------------------------	------------

Tool/service

Tool/service type	Tool	
Language dependent	False	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	Linux	
Tool/service creation	Implementation language	C++

1.11. Hungarian Opinion-Tagged Sentence Bank

General Information

Short name	OpinHuBank

Description	The OpinHuBank is a human-annotated resource for researching, evaluating and developing opinion mining systems for Hungarian. The resource consists of several thousand sentences selected from Hungarian online newswire, blogs and social media. Named entities are identified in each sentence with automatic NER tools. 5 independent human annotators were asked to indicate what polarity (opinion) was expressed towards each entity in each sentence (neutral, positive or negative).
Identifier	111
Resource type	Corpus
URL	http://www.nytud.hu/depts/corpus/index.html
Version	1.0
Revision	compilation of the corpus
Last update	2012-06-30

Contacts

Tibor Pinter	
Position	research fellow
Contact	Benczur utca 33. 1068 Budapest pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
Organization	Research Institute for Linguistics, Hungarian academy of Sciences Department of Language Technology linguinst@nytud.mta.hu
Tamás Prajczér	
Position	CEO
Contact	Bécsi út 126-128. 1034 Budapest prajczér@geox.hu http://geox.hu/ceginformacio/kapcsolat/prajczér
Organization	GeoX Térinformatikai Kft. prajczér@geox.hu

Distribution

Availability	Available – unrestricted use	
IPR holder	GeoX Térinformatikai Kft.	
	Short name	GeoX
	Contact	Bécsi út 126-128. 1034 Budapest prajczér@geox.hu

	http://geox.hu/
Availability start date	2012-07-01

Licences

CC-BY		
Restrictions of use	Academic – non-commercial use Attribution Commercial use	
Access medium	Hard disk	
Signatories	Tamás Prajczér	
	Position	CEO
	Contact	Bécsi út 126-128. 1034 Budapest prajczér@geox.hu http://geox.hu/ceginformacio/kapcsolat/prajczér
	Organization	GeoX Térinformatikai Kft. prajczér@geox.hu
Distribution rights holder	GeoX Térinformatikai Kft.	
	Short name	GeoX
	Contact	Bécsi út 126-128. 1034 Budapest prajczér@geox.hu http://geox.hu/

Metadata

Creation date	2012-06-25	
Metadata creators	Tibor Pinter	
	Position	research fellow
	Contact	Benczur utca 33. 1068 Budapest pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
	Organization	Research Institute for Linguistics, Hungarian academy of Sciences Department of Language Technology linguinst@nytud.mta.hu
Source	CESAR	
Metadata language ID	en	

Metadata last date updated	2012-06-25
-----------------------------------	------------

Texts

Media type	text	
Linguality type	Monolingual	
Multilinguality type	Parallel	
Languages	Hungarian	
	Language ID	hu
	Size	10006
	Hungarian	
	Language ID	hu
	Size	8145 sentences
Modality	Modality type	Other
Size	8154 sentences	
Text format	text/csv	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Sentence
	Format	text/csv
	Conformance to standards best practices	Other
	Annotation mode	Mixed
	Annotation tool	http://mokk.bme.hu/resources/huntoken/
	Size	204361 words
	Annotators	
	Anna Zsíros	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft. prajczer@geox.hu
	Kornél Koósz	
	Contact	Bécsi út 126-128.

	1034 Budapest prajczer@geox.hu http://geox.hu/
Organization	GeoX Térinformatikai Kft. prajczer@geox.hu
Júlia Domán	
Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
Organization	GeoX Térinformatikai Kft. prajczer@geox.hu
Andrea Koronkai	
Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
Organization	GeoX Térinformatikai Kft. prajczer@geox.hu
Zsófia Csikár	
Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
Organization	GeoX Térinformatikai Kft. prajczer@geox.hu
Other	
Segmentation level	Word
Format	text/csv
Conformance to standards best practices	Other
Annotation mode	Automatic
Annotation tool	http://mokk.bme.hu/resources/huntag/
Size	10006
Annotators	Anna Zsíros
	Contact Bécsi út 126-128. 1034 Budapest

		prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft. prajczer@geox.hu
	Kornél Koós	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft. prajczer@geox.hu
	Júlia Domán	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft. prajczer@geox.hu
	Andrea Koronkai	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft. prajczer@geox.hu
	Zsófia Csikár	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft. prajczer@geox.hu
	Segmentation	
	Segmentation level	Sentence
	Format	text/csv
	Conformance to standards best practices	Other
	Annotation	Mixed

mode		
Annotation tool	http://mokk.bme.hu/resources/huntoken/	
Size	204361 words	
Annotators	Anna Zsíros	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai K.ft. prajczer@geox.hu
	Kornél Koósz	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai K.ft. prajczer@geox.hu
	Júlia Domán	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai K.ft. prajczer@geox.hu
	Andrea Koronkai	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai K.ft. prajczer@geox.hu
	Zsófia Csikár	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai K.ft. prajczer@geox.hu

1.12. HunNERwiki: Automatically generated NE tagged corpus for Hungarian

General Information

Short name	hunNERwiki
Description	The text of the corpus is automatically generated from Hungarian Wikipedia articles. It contains Named Entity (NE) tagging according to the CoNLL standard (Person, Organization, Location and Miscellaneous), and additional morphological annotation. The corpus is the largest ever NE tagged corpus for Hungarian, which can be used for training and testing NE recognizer applications. Thanks to the standard tagset, the performance of systems trained on the hunNERwiki corpus is comparable with the performance of other state-of-the-art systems. Besides the obvious advantages of fully automatic building and annotation procedure (reducing the annotation cost), the novelty of the corpus is the application of collaboratively constructed resources (Wikipedia, DBpedia).
Identifier	112
Resource type	Corpus
URL	http://hlt.sztaki.hu/resources/hunnerwiki.html
Version	1.0
Revision	compilation of the corpus
Last update	2012-06-30

Contacts

Dávid Márk Nemeskey	
Position	research associate
Contact	Lágymányosi u. 11. 1111 Budapest nemeskey.david@sztaki.hu http://www.sztaki.hu/munkatars/008007760/
Organization	Computer and Automation Research Institute, Hungarian Academy of Sciences nemeskey.david@sztaki.hu
Tibor Pinter	
Position	research fellow
Contact	Benczur utca 33. 1068 Budapest pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
Organization	Research Institute for Linguistics, Hungarian academy of Sciences Department of Language Technology linguinst@nytud.mta.hu

Distribution

Availability	Available – unrestricted use	
IPR holder	Computer and Automation Research Institute, Hungarian Academy of Sciences	
	Short name	MTA SZTAKI
	Contact	Lágymányosi u. 11. 1111 Budapest nemeskey.david@sztaki.hu http://www.sztaki.hu/
Availability start date	2012-07-01	

Licences

CC-BY-SA		
Restrictions of use	Attribution	
	Share alike	
Access medium	Hard disk	
Signatories	Dávid Márk Nemeskey	
	Position	research associate
	Contact	Lágymányosi u. 11. 1111 Budapest nemeskey.david@sztaki.hu http://www.sztaki.hu/munkatars/008007760/
	Organization	Computer and Automation Research Institute, Hungarian Academy of Sciences nemeskey.david@sztaki.hu
Distribution rights holder	Computer and Automation Research Institute, Hungarian Academy of Sciences	
	Short name	MTA SZTAKI
	Contact	Lágymányosi u. 11. 1111 Budapest nemeskey.david@sztaki.hu http://www.sztaki.hu/

Metadata

Creation date	2012-06-25	
Metadata creators	Tibor Pinter	
	Position	research fellow
	Contact	Benczur utca 33. 1068 Budapest

		pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
	Organization	Research Institute for Linguistics, Hungarian academy of Sciences Department of Language Technology linguinst@nytud.mta.hu
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-06-25	

Usage

Actual uses	NLP applications	
	Reports	Eszter Simon, Dávid M. Nemeskey. 2012. Automatically generated NE tagged corpora for English and Hungarian. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 38-46.

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	hu
	Size	19108597 words
Modality	Modality type	Written language
Size	19108597 words	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Semantic annotation – named entities	
	Segmentation level	Word
	Format	text/csv
	Conformance to standards best practices	Other
	Annotation mode	Automatic
	Annotation tool	in-house software, hunmorph, hundiambig

	Size	19108597 words
	Annotators	Dávid Márk Nemeskey
		Position research associate
		Contact Lágymányosi u. 11. 1111 Budapest nemeskey.david@sztaki.hu http://www.sztaki.hu/munkatars/008007760/
		Organization Computer and Automation Research Institute, Hungarian Academy of Sciences nemeskey.david@sztaki.hu
		Eszter Simon
		Position research fellow
		Contact Benczúr utca 33. 1068 Budapest eszter@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Simon_Eszter.html
		Organization Research Institute for Linguistics, Hungarian Academy of Sciences Department of Language Technology pinter.tibor@nytud.mta.hu

1.13. Hungarian Verb Phrase Constructions

General Information

Short name	HVPC
Description	Hungarian Verb Phrase Constructions is a list of verb phrase constructions (VPC) automatically extracted from the Hungarian National Corpus. VPCs consist of a verb and zero or more noun phrases or postpositional phrases either lexically fixed or lexically free (cf. the English VPC 'to take sth into consideration' has a lexically free direct object and a lexically fixed into-PP). The resource also contains frequency information.
Identifier	113
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Computational lexicon
URL	http://www.nytud.hu/depts/corpus/index.html

Contacts

Bálint Sass	
Position	research fellow
Contact	Benczúr utca 33.

	1068 Budapest sass.balint@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Sass_Balint.html
Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Language Technology linguinst@nytud.mta.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

CC-BY-NC-SA	
Restrictions of use	Academic – non-commercial use
Access medium	Downloadable

Metadata

Creation date	2012-06-25	
Metadata creators	Tibor Pinter	
	Position	research fellow
	Contact	Benczur utca 33. 1068 Budapest pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
	Organization	Research Institute for Linguistics, Hungarian academy of Sciences Department of Language Technology pinter.tibor@nytud.mta.hu
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-06-25	

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds Own funds

	Funder	European Commission (50%) Research Institute for Linguistics, Hungarian academy of Sciences (50%)
	Country	Hungary
	Start date	2011-02-01
	End date	2013-01-31

Lexical conceptual resource

Lexical conceptual resource type	Computational lexicon
---	-----------------------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	hu
	Language script	latin
Modality	Modality type	Written language
Size	6266 units	
Text format	text	
Character encoding	UTF-8	

1.14. hunner

General Information

Short name	hunner
Description	Huntag can perform any kind of supervised sequential sentence tagging tasks. It has been used for NP chunking, Named Entity Recognition, and clause chunking. The flexibility of Huntag comes from the fact that it will generate any kind of features from the input data given the appropriate python functions. Several dozens of features used regularly in NLP tasks are already implemented in the file features.py, however the user is encouraged to add any number of her own. Once the desired features are implemented, a data set and a configuration file containing the list of feature functions to be used are all Huntag needs to perform training and tagging. hunner is huntag's instantiation for Named Entity Recognition.
Identifier	114
Resource type	Tool/service
Tool/service type	Tool
URL	http://mokk.bme.hu/resources/huntag

Contacts

Dániel Varga	
Position	assistant researcher
Contact	Stoczek utca 2. 1111 Budapest daniel@mokk.bme.hu http://szoc.bme.hu/oktatok/varga_daniel
Organization	Budapest University of Technology and Economics MOKK Centre for Media Research and Education daniel@mokk.bme.hu

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

LGPL	
Restrictions of use	Share alike
Access medium	Downloadable

Metadata

Creation date	2012-07-02	
Metadata creators	Tibor Pinter	
	Position	research fellow
	Contact	Benczur utca 33. 1068 Budapest pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
	Organization	Research Institute for Linguistics, Hungarian academy of Sciences Department of Language Technology pinter.tibor@nytud.mta.hu
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-02	

Usage

Actual uses	NLP applications
--------------------	-------------------------

	Reports	Gábor Reeski, Dániel Varga 2010. A Hungarian NP-chunker. In: Márton Sóskuthy (ed.) The Odd Yearbook, Budapest. p. 87-93.
--	----------------	--

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds Own funds
	Funder	European Commission (50%) Research Institute for Linguistics, Hungarian academy of Sciences (50%)
	Country	Hungary
	Start date	2011-02-01
	End date	2013-01-31

Tool/service

Tool/service type	Tool	
Language dependent	False	
Input	Media type	text
	Modality type	Written language

1.15. hunpars

General Information

Short name	hunpars
Description	hunpars is a syntactic analyzer developed for Hungarian language. hunpars can explore the syntactic structure of the simple Hungarian sentences. The elements of the grammatical hierarchy of sentences made by this syntactical analyzer are tagged by morphological features. The application is developed on rule-based system.
Identifier	115
Resource type	Tool/service
Tool/service type	Tool
URL	http://mokk.bme.hu/resources/hunpars

Contacts

--

Dániel Varga	
Position	assistant researcher
Contact	Stoczek utca 2. 1111 Budapest daniel@mokk.bme.hu http://szoc.bme.hu/oktatok/varga_daniel
Organization	Budapest University of Technology and Economics MOKK Centre for Media Research and Education daniel@mokk.bme.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

LGPL	
Restrictions of use	Share alike
Access medium	Downloadable

Metadata

Creation date	2012-07-02	
Metadata creators	Tibor Pinter	
	Position	research fellow
	Contact	Benczur utca 33. 1068 Budapest pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
	Organization	Research Institute for Linguistics, Hungarian academy of Sciences Department of Language Technology pinter.tibor@nytud.mta.hu
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-02	

Usage

Actual uses	NLP applications	
	Reports	Babarczy Anna – Gábor B. – Hamp Gábor – Kárpáti A. – Rung András – Szakadát István 2005. Mondattani elemző alkalmazás, In:

	Alexin Zoltán – Csendes Dóra (szerk.), III. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, 20–28.
--	---

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds Own funds
	Funder	European Commission (50%) Research Institute for Linguistics, Hungarian academy of Sciences (50%)
	Country	Hungary
	Start date	2011-02-01
	End date	2013-01-31

Tool/service

Tool/service type	Tool	
Language dependent	False	
Input	Media type	text
	Modality type	Written language

1.16. hunpos

General Information

Short name	hunpos
Description	hunpos follows the architecture of Thorsten Brandt's TnT system (tag n-gram HMM, with the suffix guessing emission model for unseen words), but with the ability to incorporate the output of a morphological analyzer, using the output of the morphological analyzer to constrain suffix guessing when meeting unseen words. As an improved, open source reimplementation of TnT, it is frequently used for the task of POS-tagging and morphological disambiguation, and is one of the standard building blocks and baselines when creating new POS-tagging systems and evaluating their precision.
Identifier	116
Resource type	Tool/service
Tool/service type	Tool
URL	http://mokk.bme.hu/resources/hunpos

Contacts

Dániel Varga	
Position	assistant researcher
Contact	Stoczek utca 2. 1111 Budapest daniel@mokk.bme.hu http://szoc.bme.hu/oktatok/varga_daniel
Organization	Budapest University of Technology and Economics MOKK Centre for Media Research and Education daniel@mokk.bme.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

LGPL	
Restrictions of use	Share alike
Access medium	Downloadable

Metadata

Creation date	2012-07-02	
Metadata creators	Tibor Pinter	
	Position	research fellow
	Contact	Benczur utca 33. 1068 Budapest pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
	Organization	Research Institute for Linguistics, Hungarian academy of Sciences Department of Language Technology pinter.tibor@nytud.mta.hu
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-02	

Usage

Actual uses	NLP applications
--------------------	-------------------------

	Reports	Alácsy Péter, Kornai András, Oravecz Csaba 2007. HunPos: an open source trigram tagger. ANNUAL MEETING- ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 2007, CONF 45; VOL 2, pages 2-209-2-212. http://acl.ldc.upenn.edu/P/P07/P07-2053.pdf
--	----------------	--

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds Own funds
	Funder	European Commission (50%) Research Institute for Linguistics, Hungarian academy of Sciences (50%)
	Country	Hungary
	Start date	2011-02-01
	End date	2013-01-31

Tool/service

Tool/service type	Tool	
Language dependent	False	
Input	Media type	text
	Modality type	Written language

1.17. Hungarian WSD Corpus

General Information

Short name	HunWSD
Description	The Hungarian WSD corpus contains 300-500 occurrences of 39 word forms that were selected for the purpose of word sense disambiguation. The Hungarian National Corpus and its Heti Világgazdaság (HVG) subcorpus provided the basis for corpus text selection. Texts were annotated by two independent annotators and differences were disambiguated by a third one.
Identifier	117
Resource type	Corpus
URL	http://www.inf.u-szeged.hu/rgai/corpus_hunwsl

Contacts

Veronika Vincze	
Position	research fellow
Contact	Tisza Lajos krt. 103. 6720 Szeged vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html
Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu
Richárd Farkas	
Contact	Tisza Lajos krt. 103. 6720 Szeged rfarkas@inf.u-szeged.hu http://www.inf.u-szeged.hu/~rfarkas
Organization	University of Szeged Department of Informatics rfarkas@inf.u-szeged.hu
János Csirik	
Contact	Árpád tér 2. 6720 Szeged csirik@inf.u-szeged.hu http://www.inf.u-szeged.hu/~csirik
Organization	University of Szeged Department of Informatics csirik@inf.u-szeged.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

MS-NC-NoReD		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://www.inf.u-szeged.hu/rgai/corpus_hunwsd	
Signatories	Veronika Vincze	
	Position	research fellow
	Contact	Tisza Lajos krt. 103. 6720 Szeged vinczev@inf.u-szeged.hu

		http://www.inf.u-szeged.hu/~vinczev/index_en.html
	Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu
Distribution rights holder	University of Szeged	
	Short name	SZTE
	Department name	Department of Informatics
	Contact	Tisza Lajos krt. 103. 6720 Szeged vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html

Metadata

Creation date	2012-06-26	
Metadata creators	Veronika Vincze	
	Contact	vinczev@inf.u-szeged.hu
	Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu

Usage

Actual uses	NLP applications	
	Reports	Vincze, Veronika; Szarvas, György; Almási, Attila; Szauter, Dóra; Ormándi, Róbert; Farkas, Richárd; Hatvani, Csaba; Csirik, János 2008: Hungarian Word-sense Disambiguated Corpus. In: Proceedings of 6th International Conference on Language Resources and Evaluation LREC 2008, Marrakech, Morocco.

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01

	End date	2013-01-31
--	-----------------	------------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	14000
Modality	Modality type	Other
Size	14000	
Character encoding	UTF-8	
Annotation	Semantic annotation	
	Segmentation level	Word

1.18. Hungarian Language Processing Tools in NooJ

General Information

Short name	NooJ
Description	<p>The Hungarian NooJ contains a morphological dictionary (based on the more than 60 000 lemmata found in the Concise Dictionary of Hungarian Language morphological information based on the work of Laszlo Elekfi). From the base forms and the morphological information contained in the .DIC files using the inflectional rules described in the .FLX files complex inflected forms of nouns and verbs are generated with the help of NooJ compile dictionary function. The result of the compilation can be found in the .NOD files. With the aid of the NOD files complex inflected forms can be recognised in the running texts, including derived and further inflected running words, as well as non inflected forms, naturally. Separate dictionaries contain words which cannot be inflected. As the result of this, complex suffixed words and/or compounds can also be recognised when analysing a text. With the aid of the compiled dictionaries and the language specific syntactic graphs the tool performs sentence- and clause-segmentation, POS-tagging NP-recognition, predicate-identification and the identification of the other sentence constituents (eg. adverbials). The input text may be any Hungarian raw text or any xml-text compatible with NooJ, and the output may also be exported in xml-format. NooJ is widely used in Hungarian linguistics and language technology: its usage covers a broad scale of morphological, syntactic, lexical, semantic and psychological content analyses. The Hungarian NooJ tools are consisting of a range of specific dictionaries (basic .dic files for dictionaries, .nog files for compiled dictionaries and .flx files for morphological rules). Each of them is created for specific analyses. Below is a short description for each of them: noun.dic Hungarian nouns supplied with morphological information -- 55000 units, verb_00.dic Hungarian verbs supplied with morphological information -- 10000 units, topabbr.dic Most frequent Hungarian abbreviations -- 11 tokens, noaffix-nins.dic Hungarian words which cannot be inflected -- 1870 units, topprop.dic Most frequent proper names -- 28 units, noun.nod Compiled NooJ dictionary of Hungarian nouns -- 96777513 words, verb_00.nod Compiled NooJ dictionary of Hungarian verbs -- 19059644, topabbr.nod Most frequent Hungarian</p>

	abbreviations -- 11 words, noaffix.nod Compiled Nooj dictionary of Hungarian words which cannot be inflected. -- 1870 words, topprop.nod Compiled Nooj dictionary of the most frequent Hungarian proper names -- 28 words, noun.flx Inflectional rules of Hungarian nouns according to their morphological category -- 33 000 rules, verb.flx Inflectional rules of Hungarian verbs according to their morphological category -- 27900 rules
Identifier	118
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Computational lexicon
URL	http://corpus.nytud.hu/nooj

Contacts

Júlia Pajzs	
Position	senior research fellow
Contact	Benczur utca 33. 1068 Budapest pajzs.julia@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pajzs_Julia.html
Organization	Research Institute for Linguistics, Hungarian academy of Sciences Department of Language Technology pinter.tibor@nytud.mta.hu

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

GPL	
Restrictions of use	Share alike
Access medium	Downloadable

Metadata

Creation date	2012-06-25	
Metadata creators	Tibor Pinter	
	Position	research fellow
	Contact	Benczur utca 33. 1068 Budapest pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
	Organization	Research Institute for Linguistics, Hungarian academy of Sciences

	Department of Language Technology pinter.tibor@nytud.mta.hu
Source	CESAR
Metadata language ID	en
Metadata last date updated	2012-06-25

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds Own funds
	Funder	European Commission (50%) Research Institute for Linguistics, Hungarian academy of Sciences (50%)
	Country	Hungary
	Start date	2011-02-01
	End date	2013-01-31

Lexical conceptual resource

Lexical conceptual resource type	Computational lexicon
---	-----------------------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	hu
	Language script	latin
Modality	Modality type	Written language
Size	see: description tokens	
Text format	text	
Character encoding	UTF-8	

1.19. SzegedParalell

General Information

Short name	SzegedParalell
Description	The English-Hungarian parallel corpus contains texts selected on the basis of grammatical and translational criteria. Sentences representing the grammar of the given language (usually taken from language books) and authentic texts are both included in the parallel corpus, thus, the balance is maintained between artificially constructed and natural language structures. Both paragraph and sentence alignment were checked and corrected manually.
Identifier	119
Resource type	Corpus
URL	http://www.inf.u-szeged.hu/rgai/corpus_paralell

Contacts

Richard Farkas	
Contact	Tisza Lajos krt. 103. 6720 Szeged rfarkas@inf.u-szeged.hu http://www.inf.u-szeged.hu/~rfarkas
Organization	University of Szeged Department of Informatics rfarkas@inf.u-szeged.hu
Janos Csirik	
Contact	Arpad ter 2. 6720 Szeged csirik@inf.u-szeged.hu http://www.inf.u-szeged.hu/~csirik
Organization	University of Szeged Department of Informatics csirik@inf.u-szeged.hu
Veronika Vincze	
Position	research fellow
Contact	Tisza Lajos krt. 103. 6720 Szeged vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html
Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

MS-NC-NoReD	
Restrictions of use	Academic – non-commercial use
Access medium	Downloadable
Download location	http://www.inf.u-szeged.hu/rgai/corpus_paralell
Signatories	Veronika Vincze
	Position research fellow
	Contact Tisza Lajos krt. 103. 6720 Szeged vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html
	Organization University of Szeged Department of Informatics vinczev@inf.u-szeged.hu
Distribution rights holder	University of Szeged
	Short name SZTE
	Department name Department of Informatics
	Contact Tisza Lajos krt. 103. 6720 Szeged vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html

Metadata

Creation date	2012-06-26
Metadata creators	Veronika Vincze
	Contact vinczev@inf.u-szeged.hu
	Organization University of Szeged Department of Informatics vinczev@inf.u-szeged.hu

Usage

Actual uses	NLP applications	
	Reports	Toth, Krisztina; Farkas, Richard; Kocsor, András 2008: Sentence alignment of Hungarian-English parallel corpora using a hybrid

	algorithm. ACTA Cybernetica 18:463-478.
--	---

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2013-01-31

Texts

Media type	text	
Linguality type	Bilingual	
Languages	Hungarian	
	Language ID	HU
	Size	99000 sentences
	English	
	Language ID	EN
	Size	99000 sentences
Size	99000 sentences	
Annotation	Alignment	

1.20. Szeged Criminal NE Corpus

General Information

Short name	SzegedCriNE
Description	The corpus contains texts on criminal offences which are annotated for named entites.
Identifier	120
Resource type	Corpus
URL	http://www.inf.u-szeged.hu/rgai/corpus_ne

Contacts

--

Richard Farkas	
Contact	Tisza Lajos krt. 103. 6720 Szeged rfarkas@inf.u-szeged.hu http://www.inf.u-szeged.hu/~rfarkas
Organization	University of Szeged Department of Informatics rfarkas@inf.u-szeged.hu
Janos Csirik	
Contact	Arpad ter 2. 6720 Szeged csirik@inf.u-szeged.hu http://www.inf.u-szeged.hu/~csirik
Organization	University of Szeged Department of Informatics csirik@inf.u-szeged.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

MS-NC-NoReD	
Restrictions of use	Academic – non-commercial use
Access medium	Downloadable
Download location	http://www.inf.u-szeged.hu/rgai/corpus_ne
Signatories	Veronika Vincze
	Position research fellow
	Contact Tisza Lajos krt. 103. 6720 Szeged vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html
	Organization University of Szeged Department of Informatics vinczev@inf.u-szeged.hu
Distribution rights holder	University of Szeged
	Short name SZTE
	Department name Department of Informatics
	Contact Tisza Lajos krt. 103.

		6720 Szeged vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html
--	--	--

Metadata

Creation date	2012-06-26	
Metadata creators	Veronika Vincze	
	Contact	vinczev@inf.u-szeged.hu
	Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2013-01-31

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	540000 tokens
Modality	Modality type	Other
Size	540000 tokens	
Annotation	Semantic annotation – named entities	

1.21. SzegedParallelFX

General Information

Short name	SzPFX
Description	The SzegedParalell corpus constitutes the basis of the SzegedParalellFX, in which light verb constructions are annotated (14,261 sentence alignment units in size containing 1370 occurrences of light verb constructions).
Identifier	121
Resource type	Corpus
URL	http://www.inf.u-szeged.hu/rgai/mwe

Contacts

Richard Farkas	
Contact	Tisza Lajos krt. 103. 6720 Szeged rfarkas@inf.u-szeged.hu http://www.inf.u-szeged.hu/~rfarkas
Organization	University of Szeged Department of Informatics rfarkas@inf.u-szeged.hu
Janos Csirik	
Contact	Arpad ter 2. 6720 Szeged csirik@inf.u-szeged.hu http://www.inf.u-szeged.hu/~csirik
Organization	University of Szeged Department of Informatics csirik@inf.u-szeged.hu
Veronika Vincze	
Position	research fellow
Contact	Tisza Lajos krt. 103. 6720 Szeged vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html
Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

--

MS-NC-NoReD		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://www.inf.u-szeged.hu/rgai/mwe	
Signatories	Veronika Vincze	
	Position	research fellow
	Contact	Tisza Lajos krt. 103. 6720 Szeged vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html
	Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu
Distribution rights holder	University of Szeged	
	Short name	SZTE
	Department name	Department of Informatics
	Contact	Tisza Lajos krt. 103. 6720 Szeged vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html

Metadata

Creation date	2012-06-26	
Metadata creators	Veronika Vincze	
	Contact	vinczev@inf.u-szeged.hu
	Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu

Usage

Actual uses	NLP applications	
	Reports	Vincze, Veronika 2012: Light Verb Constructions in the SzegedParallelFX English-Hungarian Parallel Corpus. In: Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC 2012). Istanbul, Turkey, pp. 2381-2388.

Resource creation

Funding projects	Central and South-East European Resources	

	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2013-01-31

Texts

Media type	text	
Linguality type	Bilingual	
Languages	Hungarian	
	Language ID	HU
	Size	14000 sentences
	English	
	Language ID	EN
	Size	14000 sentences
Size	14000 sentences	
Annotation	Other	

1.22. Szeged Treebank FX

General Information

Short name	Szeged Treebank FX
Description	The Szeged Treebank was annotated for light verb constructions manually. This version contains 6734 occurrences of 1215 light verb constructions altogether in 82,099 sentences.
Identifier	122
Resource type	Corpus
URL	http://www.inf.u-szeged.hu/rgai/mwe

Contacts

Veronika Vincze	
Position	research fellow
Contact	Tisza Lajos krt. 103. 6720 Szeged

	vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html
Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

MS-NC-NoReD		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://www.inf.u-szeged.hu/rgai/mwe	
Signatories	Veronika Vincze	
	Position	research fellow
	Contact	Tisza Lajos krt. 103. 6720 Szeged vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html
	Organization	University of Szeged Department of Informatics vinczev@inf.u-szeged.hu
Distribution rights holder	University of Szeged	
	Short name	SZTE
	Department name	Department of Informatics
	Contact	Tisza Lajos krt. 103. 6720 Szeged vinczev@inf.u-szeged.hu http://www.inf.u-szeged.hu/~vinczev/index_en.html

Metadata

Creation date	2012-06-26	
Metadata creators	Veronika Vincze	
	Contact	vinczev@inf.u-szeged.hu
	Organization	University of Szeged Department of Informatics

	vinczev@inf.u-szeged.hu
--	--

Usage

Actual uses	NLP applications	
	Reports	Vincze, Veronika; Csirik, Janos 2010: Hungarian Corpus of Light Verb Constructions. In: Proceedings of COLING 2010, Beijing, China, pp. 1110-1118.

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2013-01-31

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Modality	Modality type	Other
Size	82000 sentences	
Annotation	Other	

1.23. Hungarian National Corpus

General Information

Short name	HNC
Description	The national corpus of Hungarian language which is derived into five subcorpora by regional language variants, and into five subcorpora by text genres also. The subcorpus to be studied can be chosen by any combination of these. That makes the HNC an appropriate tool to study the differences not just between text genres but between language variants. HNC wishes to be a representative general-aim corpus of present-day standard

	Hungarian.HNC v2 is based on the Hungarian National Corpus with higher quality and finer level of analysis and annotation (detailed morphosyntactic analysis and disambiguation with updated processing toolchain, NP chunking, Named Entity recognition, distributional analysis, built in post-processing (multilevel frequency lists, subsequent searches on previous results)). HNC2 is extended up to 1 gigaword threshold with extended metadata and cleared IPR.
Identifier	123
Resource type	Corpus
URL	http://corpus.nytud.hu

Contacts

Casba Oravecz	
Contact	Benczúr utca 33. 1068 Budapest oravecz.csaba@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Oravecz_Csaba.html

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

MSCommons-BY-NC	
Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface

Metadata

Creation date	2013-01-25
----------------------	------------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	100 976 483 tokens
Modality	Modality type	Other
Size	100 976 483 tokens	
Annotation	Morphosyntactic annotation - below POS tagging	

1.24. HuComTech Multimodal Corpus and Database

General Information

Short name	HuComTech
Description	The HuComTech multimodal corpus consists of about 50 hours of video and audio recordings of 111 formal dialogues (simulated job interviews) and 111 informal but guided dialogues. The language of the recordings is Hungarian. The participants were university students aged 19-27, female 54 and male 67. The corpus was annotated for video (facial expressions, instances of eyebrows, gaze, headshift, handshape, touchmotion and posture) and audio (emotions, discourse, prosody and textual transcriptions). Its unique features in a wider comparison include its special attention to pragmatics focusing on a comparative study of the unimodal vs. multimodal features of communication (as compared to multimodality alone) as well as the study of the syntax and prosody of spoken language with respect to the above wide range of multimodal characteristics. The data can be queried in ELAN and in our web-based SQL database.
Identifier	124
Resource type	Corpus
URL	https://hucomtech.unideb.hu/hucomtech/

Contacts

László Hunyadi	
Position	research manager
Contact	Egyetem tér 1. 4032 Debrecen hunyadi@ling.arts.unideb.hu http://lingua.arts.unideb.hu
Organization	University of Debrecen Department of General and Applied Linguistics hunyadi@ling.arts.unideb.hu
László Hunyadi	
Position	research manager
Contact	Egyetem tér 1. 4032 Debrecen hunyadi@ling.arts.unideb.hu http://www.lingua.arts.unideb.hu
Organization	University of Debrecen Department of General and Applied Linguistics hunyadi@ling.arts.unideb.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

IPR holder	Department of General and Applied Linguistics, University of Debrecen	
	Contact	Egyetem tér 1. 4032 Debrecen hunyadi@ling.arts.unideb.hu http://lingua.arts.unideb.hu
Availability start date	2013-01-29	

Licences

MS-NC-NoReD		
Restrictions of use	Academic – non-commercial use No derivatives	
Access medium	Web executable	
Signatories	László Hunyadi	
	Position	research manager
	Contact	Egyetem tér 1. 4032 Debrecen hunyadi@ling.arts.unideb.hu
	Organization	University of Debrecen Department of General and Applied Linguistics hunyadi@ling.arts.unideb.hu
Distribution rights holder	Department of General and Applied Linguistics, University of Debrecen	
	Contact	Egyetem tér 1. 4032 Debrecen hunyadi@ling.arts.unideb.hu http://lingua.arts.unideb.hu

Metadata

Creation date	2013-01-08	
Metadata creators	István Szekrényes	
	Position	research assistant
	Contact	Egyetem tér 1. 4032 Debrecen xepenator@gmail.com http://www.lingua.arts.unideb.hu
	Organization	University of Debrecen Department of General and Applied Linguistics hunyadi@ling.arts.unideb.hu
Source	CESAR	

Metadata language ID	en
Metadata last date updated	2013-01-08

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	hu
Modality	Modality type	Spoken language
Audio size	50 hours	

1.25. BEA Hungarian spontaneous speech database

General Information

Short name	BEA
Description	<p>The aim of developing a phonetically-based multi-purpose database of Hungarian spontaneous speech, dubbed BEA (BESzél nyelv Adatbázis ‘spoken language database’), is to accumulate a large amount of recorded spontaneous speech produced by numerous present-day Budapest speakers, providing ample material for various types of research and practical applications. At the time of writing, the total recorded material of BEA is 250 hours long, meaning approximately 3,400,000 running words. The database primarily contains spontaneous speech materials, but for the sake of comparisons, it also includes sentence repetitions and read texts. The database offers material for research in a number of areas within linguistics. The study of acoustic-phonetic consequences of the production of speech sounds, coarticulation effects, and suprasegmental features was hampered for many years by the methodological difficulty that no spontaneous speech material of an adequate quality and quantity was available. The database contains transcriptions of the BEA speech materials at several levels: i) primary transcription in orthography but without punctuation. Transcribers use Microsoft Office Word (.doc format); ii) annotation: This form of transcription is a kind of visual display of spoken texts and some further pieces of information related to them in a way that the written text and the actual recording can be displayed/listened to simultaneously. This is made possible by software Transcriber. In addition to phonetic research in the strict sense, now it becomes possible to carry on conversation analysis, pragmatic research, speech technology, the study of speech accommodation, that of the spontaneous speech of elderly speakers, or that of disfluency phenomena.</p>
Identifier	125
Resource type	Corpus
URL	http://www.nytud.hu/adatb/bea/index.html
Version	1.0
Last update	2013-01-31

Contacts

Dorottya Gyarmathy	
Position	research fellow
Contact	Benczúr u. 33. 1068 Budapest gyarmathy.dorottya@nytud.mta.hu http://www.nytud.hu/depts/phonetics/gyarmathyd/index.html
Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Phonetics linginst@nytud.mta.hu
András Beke	
Position	research fellow
Contact	Benczúr utca 33. 1068 Budapest beke.andras@nytud.mta.hu http://www.nytud.hu/depts/phonetics/beke/index.html
Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Phonetics linguinst@nytud.mta.hu

Distribution

Availability	Available – restricted use	
IPR holder	Research Institute for Linguistics, Hungarian Academy of Sciences	
	Short name	RIL HAS
	Contact	Benczúr u. 33. 1068 Budapest linginst@nytud.mta.hu http://www.nytud.hu/eng/index.html
Availability start date	2013-01-29	

Licences

MS-NC-NoReD-ND		
Restrictions of use	Academic – non-commercial use	
	No derivatives	
Access medium	Downloadable	
Signatories	Mária Gósy	
	Position	research fellow

	Contact	Benczúr u. 33. 1068 Budapest gosity.maria@nytud.mta.hu http://www.nytud.hu/depts/phonetics/gosity/index.html
	Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Phonetics linginst@nytud.mta.hu
Distribution rights holder	Research Institute for Linguistics, Hungarian Academy of Sciences	
	Short name	RIL MTA
	Contact	Benczúr u. 33. 1068 Budapest linginst@nytud.mta.hu http://www.nytud.hu/eng/index.html

Metadata

Creation date	2013-01-08	
Metadata creators	Tibor Pintér	
	Position	research fellow
	Contact	Benczúr utca 33. 1068 Budapest pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
	Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Language Technology linguinst@nytud.mta.hu
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-08	

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission

	Country	European Union
	Start date	2011-02-01
	End date	2013-01-31

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	hu
	Size	3400000 words
Modality	Modality type	Spoken language
Size	3400000 words	
Text format	text	
Character encoding	UTF-8	

1.26. Hungarian Kindergarten Language Corpus

General Information

Short name	HUKILC
Description	The Hungarian Kindergarten Language Corpus (HUKILC) has been compiled predominantly for child language variation studies. It contains 62 interviews with 4,5-5,5 year-old kindergarten children from Budapest. The interviews are at least 20 minutes long. The children are divided into 4 groups concerning socio-economic status (SES) and sex. There is a higher SES group with males (hm), and one with females (hf), and a lower SES group with males (lm) and females (lf), respectively. The corpus is also a useful source for other fields of child language research (eg. phonetics, or developmental morphology). A morphological analyzer (Humor) and disambiguator (PurePos) has been adapted for child language data (still in progress) in HUKILC.
Identifier	126
Resource type	Corpus
URL	http://www.meta-share.nytud.hu
Version	1.0
Revision	compilation of the corpus
Last update	2013-01-21

Contacts

Kinga Mátyus	
Position	junior research fellow

Contact	Benczúr u. 33. 1068 Budapest matyus.kinga@nytud.mta.hu http://www.nytud.hu/
Organization	Research Institute for Linguistics, Hungarian Academy of Sciences matyus.kinga@nytud.mta.hu
Kinga Mátyus	
Position	junior research fellow
Contact	Benczúr u. 33. 1068 Budapest matyus.kinga@nytud.mta.hu http://www.nytud.hu/
Organization	Research Institute for Linguistics, Hungarian academy of Sciences Department of Language Technology linguinst@nytud.mta.hu

Distribution

Availability	Available – unrestricted use	
IPR holder	Research Institute for Linguistics, Hungarian Academy of Sciences	
	Short name	MTA RIL
	Contact	Benczúr u. 33. 1068 Budapest matyus.kinga@nytud.mta.hu http://www.nytud.hu/
Availability start date	2013-01-21	

Licences

MS-NC-NoReD-ND		
Restrictions of use	Academic – non-commercial use Inform licensor No derivatives Share alike	
Access medium	Downloadable	
Signatories	Kinga Mátyus	
	Position	junior research fellow
	Contact	Benczúr u. 33. 1068 Budapest matyus.kinga@nytud.mta.hu

		http://www.nytud.hu/oszt/elonyelv/mts.html
	Organization	Research Institute for Linguistics, Hungarian Academy of Sciences matyus.kinga@nytud.mta.hu
Distribution rights holder	Research Institute for Linguistics, Hungarian Academy of Sciences	
	Short name	MTA RIL
	Contact	Benczúr u. 33. 1068 Budapest matyus.kinga@nytud.mta.hu http://www.nytud.hu/

Metadata

Creation date	2013-01-21	
Metadata creators	Kinga Mátyus	
	Position	junior research fellow
	Contact	Benczúr u. 33. 1068 Budapest matyus.kinga@nytud.mta.hu http://www.nytud.hu/
	Organization	Research Institute for Linguistics, Hungarian academy of Sciences Department of Language Technology linguinst@nytud.mta.hu
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-21	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	hu
	Size	192000 words
Modality	Modality type	Spoken language
Size	192000 words	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Morphosyntactic annotation – POS tagging	

	Segmentation level	Word
	Format	text
	Conformance to standards best practices	Other
	Annotation mode	Automatic
	Annotation tool	humor, purepos
	Size	192000 words

1.27. ht-online

General Information

Short name	ht-online
Description	Ht-online is a unique lexical database of the most common loanwords in Hungarian language used outside Hungary (collected from 7 regions). The database should be used as spacial lexical resource in the Hungarian language tools based on the Hungarian morphology.
Identifier	127
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Computational lexicon
URL	http://ht.nytud.hu/htonline/

Contacts

Tibor Pintér	
Position	research fellow
Contact	Benczúr utca 33. 1068 Budapest pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Language Technology linguinst@nytud.mta.hu

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

MSCommons-BY-NC-SA	
Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface

Metadata

Creation date	2013-01-29	
Metadata creators	Tibor Pinter	
	Position	research fellow
	Contact	Benczur utca 33. 1068 Budapest pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
	Organization	Research Institute for Linguistics, Hungarian academy of Sciences Department of Language Technology pinter.tibor@nytud.mta.hu
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-29	

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds Own funds
	Funder	European Commission (50%) Research Institute for Linguistics, Hungarian academy of Sciences (50%)
	Country	Hungary
	Start date	2011-02-01
	End date	2013-01-31

Lexical conceptual resource

Lexical conceptual resource type	Computational lexicon
---	-----------------------

Texts

Media type	text		
Linguality type	Monolingual		
Languages	Hungarian		
	Language ID	hu	
	Language script	latin	
	Language variety	Language variety type	Dialect
		Language variety name	lexical elements of Hungarian spoken in Slovakia,Romania,Ukraine,Serbia,Croatia,Austria and Slovenia
Modality	Modality type	Spoken language Written language	
Size	4009 entries		
Text format	text		
Character encoding	UTF-8		

1.28. Hungarian Consize Dictionary

General Information

Short name	HCD
Description	A unique dictionary of Hungarian language of 16 000 headwords (entries) followed by frequency data. Each entry describes the most common forms (given by pragmatical reasons) of the headword. The entries are divided into meanings which counts 33 000 carefully selected and stilistically labeled meanings. The dictionary contains sentences brouhght from real language use and 3000 phrasems.
Identifier	128
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Computational lexicon
URL	http://www.meta-share.nytud.hu

Contacts

Tibor Pintér	
Position	research fellow
Contact	Benczúr utca 33. 1068 Budapest

	pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Language Technology linguinst@nytud.mta.hu

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

MSCommons-BY-NC-SA	
Restrictions of use	No derivatives No redistribution
Access medium	Downloadable

Metadata

Creation date	2013-01-29	
Metadata creators	Tibor Pinter	
	Position	research fellow
	Contact	Benczur utca 33. 1068 Budapest pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
	Organization	Research Institute for Linguistics, Hungarian academy of Sciences Department of Language Technology pinter.tibor@nytud.mta.hu
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-29	

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds Own funds

	Funder	European Commission (50%) Research Institute for Linguistics, Hungarian academy of Sciences (50%)
	Country	Hungary
	Start date	2011-02-01
	End date	2013-01-31

Lexical conceptual resource

Lexical conceptual resource type	Computational lexicon
---	-----------------------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	hu
	Language script	latin
Modality	Modality type	Spoken language Written language
Size	16000 entries	
Text format	text	
Character encoding	UTF-8	

1.29. HHC: Hungarian historical corpus

General Information

Short name	HHC
Description	Hungarian historical corpus (further as HHC) is a collection of texts written between 1772 and 1997 in different genres, containing ca. 27 million tokens. During the compilation of HHC, text samples were selected by professionals (literary historians, historians, mathematicians etc.) from printed works. A relative majority (40%) of the texts are dated from the second half of the 20th century. The corpus is the product of the Department of Lexicography and Lexicology at RIL HAS, made between 1986 and 1997, maintained continuously since then. As an innovation, genre labeling was unified. Thus, genres and text types in HHC and HNC are marked similarly, this makes possible to search data of these corpora by using the same query structure.
Identifier	129
Resource type	Corpus
URL	http://www.meta-share.nytud.hu

Version	1.0
Last update	2013-01-31

Contacts

Tibor Pintér	
Position	research fellow
Contact	Benczúr utca 33. 1068 Budapest pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Language Technology linguinst@nytud.mta.hu
Nóra	
Position	head of the department, senior research fellow
Contact	Benczúr u. 33. 1068 Budapest ittzes.nora@nytud.mta.hu http://www.nytud.hu/depts/lexi/staff.html
Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Lexicography and Lexicology linginst@nytud.mta.hu

Distribution

Availability	Available – restricted use
Availability start date	2013-01-31

Licences

MS-NC-NoReD		
Restrictions of use	Academic – non-commercial use	
	No derivatives	
Access medium	Accessible through interface	
Signatories	Nóra Ittész	
	Position	head of the department, senior research fellow
	Contact	Benczúr u. 33. 1068 Budapest ittzes.nora@nytud.mta.hu http://www.nytud.hu/depts/lexi/staff.html

	Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Lexicography and Lexicology linginst@nytud.mta.hu
--	---------------------	--

Metadata

Creation date	2013-01-08
Source	CESAR
Metadata language ID	en
Metadata last date updated	2013-01-28

Usage

Actual uses	Human use	
	Reports	Ittész, Nóra: A magyar nyelv nagyszótára. ['The Comprehensive Dictionary of Hungarian'] In: Fábán, Zsuzsanna (ed.): Szótárírás és szótárírók. ['Lexicography and lexicographers'] Budapest: Akadémiai Kiadó, 2009. pp. 65–80.

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	hu
	Size	27000000 words
Modality	Modality type	Written language
Size	27000000 words	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word
	Format	text/csv
	Conformance to standards best practices	TEI_P5
	Annotation tool	www.nytud.hu/hhc
	Size	27000000 words
	Annotators	Attila Mártonfi

		Position	research fellow
		Contact	Benczúr u. 33. 1068 Budapest martonfi.attila@nytud.mta.hu http://www.nytud.hu/oszt/lexi/martonfi/index.html
		Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Lexicography and Lexicology linguinst@nytud.mta.hu

1.30. N-grams from Hungarian National Corpus

General Information

Short name	HNCNgrams
Description	The national corpus of Hungarian language which is derived into five subcorpora by regional language variants, and into five subcorpora by text genres also. The subcorpus to be studied can be chosen by any combination of these. That makes the HNC an appropriate tool to study the differences not just between text genres but between language variants. HGC wishes to be a representative general-aim corpus of present-day standard Hungarian. HGC is based on the Hungarian National Corpus with higher quality and finer level of analysis and annotation (detailed morphosyntactic analysis and disambiguation with updated processing toolchain, NP chunking, Named Entity recognition, distributional analysis, built in post-processing (multilevel frequency lists, subsequent searches on previous results)). HGC is extended up to 1 gigaword threshold with extended metadata and cleared IPR.
Identifier	130
Resource type	Corpus
URL	http://www.meta-share.nytud.hu

Contacts

Casba Oravecz	
Contact	Benczúr utca 33. 1068 Budapest oravecz.csaba@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Oravecz_Csaba.html
Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Language Technology and Applied Linguistics linguinst@nytud.mta.hu

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

MS-NC-NoReD		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Signatories	Csaba Oravecz	
	Position	research fellow
	Contact	Benczúr utca 33. 1068 Budapest oravecz.csaba@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Oravecz_Csaba.html
	Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Language Technology and Applied Linguistics linguinst@nytud.mta.hu
Distribution rights holder	Research Institute for Linguistics, Hungarian Academy of Sciences	
	Short name	RILMTA
	Department name	Department of Language Technology and Applied Linguistics
	Contact	Benczúr utca 33. 1068 Budapest oravecz.csaba@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Oravecz_Csaba.html

Metadata

Creation date	2013-01-25	
Metadata creators	Tibor Pintér	
	Contact	pinter.tibor@nytud.mta.hu
	Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Language Technology and Applied Linguistics linguinst@nytud.mta.hu

Corpus text ngram

Media type	textNgram	
Ngram	Base item	Word
	Order	2
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU

Size	33014184 lemma bigrams, 50730372 wordform bigrams, 99457540 lemma trigrams, 124314331 wordform trigrams, 162616966 lemma 4 – grams, 177712578 wordform 4 – grams, 194983488 lemma 5 – grams, 200548340 wordform 5 – grams
-------------	---

1.31. CHSM-IC: Corpus of Hungarian School Metalanguage – Interview Corpus

General Information

Short name	CHSM-IC
Description	The present corpus contains semi-structured research interview texts recorded and transcribed for the purposes of a PhD project entitled 'Learning, following and disseminating language rules as a topic in the metalinguistic knowledge of students and their teachers' by Tamás Péter Szabó. 133 interviewees (partly students, partly teachers) were asked to speak about their linguistic routines and their opinion on various trends in language use. They were asked to evaluate linguistic trends and to explain linguistic phenomena as well. Data collection was carried out in elementary schools, training colleges and grammar schools, on years 1–4, 7 and 11 in Hungary (Budapest and ten counties). Students (years 1–4, 7 and 11) and their teacher of Hungarian grammar and literature (years 7 and 11) were interrogated. The research interviews were made with 1, 2, 3 or – in extreme cases – more interviewees. All of the interviews were recorded, transcribed and annotated by Tamás Péter Szabó.
Identifier	131
Resource type	Corpus
URL	http://www.meta-shhare.nytud.hu
Version	1.0
Last update	2013-01-31

Contacts

Tibor Pintér	
Position	research fellow
Contact	Benczúr utca 33. 1068 Budapest pinter.tibor@nytud.mta.hu http://www.nytud.hu/oszt/korpusz/Pinter_Tibor.html
Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Language Technology linguinst@nytud.mta.hu
Tamás Péter Szabó	
Position	research fellow
Contact	Benczúr u. 33. 1068 Budapest

	szabo.tamas.peter@nytud.mta.hu http://sztp.hu/index_eng.htm
Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Lexicography and Lexicology linginst@nytud.mta.hu

Distribution

Availability	Available – restricted use
Availability start date	2013-01-28

Licences

MS-NC-NoReD		
Restrictions of use	Academic – non-commercial use No derivatives	
Access medium	Hard disk	
Signatories	Tamás Péter Szabó	
	Position	research fellow
	Contact	Benczúr u. 33. 1068 Budapest szabo.tamas.peter@nytud.mta.hu http://sztp.hu/index_eng.htm
	Organization	Research Institute for Linguistics, Hungarian Academy of Sciences Department of Lexicography and Lexicology linginst@nytud.mta.hu

Metadata

Creation date	2013-01-08
Source	CESAR
Metadata language ID	en
Metadata last date updated	2013-01-08

Usage

Actual uses	Human use	
	Reports	Szabó, Tamás Péter: „Kirakunk táblákat, hogy csúnyán beszélni tilos”. A javítás mint gyakorlat és mint téma diákok és tanáraik metanyelvében [Repair as communication practice – repair as discourse topic. A multi-faceted investigation of Hungarian school metalanguage].

	Dunaszerdahely/Dunajská Streda (Slovakia): Gramma, 2012. URL: http://mek.oszk.hu/10900/10947/ Summary in English: pp. 229–234.
--	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	hu
	Size	346500 words
Modality	Modality type	Spoken language
Size	346500 words	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Semantic annotation – named entities	
	Segmentation level	Sentence
	Format	text/csv
	Conformance to standards best practices	TEI_P5
	Annotation mode	Manual
	Annotation tool	CLaRK, XMetaL
	Size	346500 words
	Annotators	Tamás Péter Szabó
		Position research fellow
		Contact Benczúr u. 33. 1068 Budapest szabo.tamas.peter@nytud.mta.hu http://sztp.hu/index_eng.htm
		Organization Research Institute for Linguistics, Hungarian Academy of Sciences Department of Lexicography and Lexicology linginst@nytud.mta.hu

2. BME-TMIT resources

2.1. Mindentudás Speech Corpus

General Information

Description	An audio collection of public lectures in Hungarian, together with transcriptions. The lectures took place as part of the Mindentudás Egyeteme television series.
Identifier	201
Resource type	Corpus

Contacts

Dániel Varga	
Contact	daniel@mokk.bme.hu
Organization	Budapest University of Technology and Economics MOKK Centre for Media Research and Education daniel@mokk.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Mindentudás Egyeteme	
	Contact	daniel@mokk.bme.hu http://mindentudas.hu/

Licences

MS-C-NoReD		
Restrictions of use	No redistribution	
Access medium	Downloadable	
Download location	ftp://mokk.bme.hu/Mindentudas/	
Signatories	Budapest University of Technology and Economics	
	Short name	BUTE
	Department name	Department of Telecommunications and Media Informatics
	Contact	Magyar tudósok körútja 2. H-1117 Budapest daniel@mokk.bme.hu http://tmit.bme.hu/
Distribution rights holder	Budapest University of Technology and Economics	
	Short name	BUTE
	Department name	Department of Telecommunications and Media Informatics
	Contact	Magyar tudósok körútja 2. H-1117 Budapest

		daniel@mokk.bme.hu http://tmit.bme.hu/
--	--	---

Metadata

Creation date	2012-10-09	
Metadata creators	Dániel Varga	
	Contact	daniel@mokk.bme.hu
	Organization	Budapest University of Technology and Economics MOKK Centre for Media Research and Education daniel@mokk.bme.hu

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Size	200 hours	
Annotation	Other	
	Segmentation level	Paragraph

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Audio size	200 hours	
Annotation	Other	
	Segmentation level	Paragraph

2.2. Word level speech database for Hungarian

General Information

Short name	Words-hu
Description	Word level speech database to study the acoustic structure of the Hungarian CV, VC, VV, VVV, CC, CCC and CCCC sound clusters. For the password to the database's zip file

	please contact us.
Identifier	202
Resource type	Corpus
URL	http://magyarbeszed.tmit.bme.hu/cvvc/index.php?hl=en
Version	1.0
Revision	Waves, texts, sound boundaries and waveform images
Last update	2012-07-09

Contacts

Gábor Olaszy	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Gábor Olaszy	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
Availability start date	2011-11-30	

Licences

CLARIN_RES	
Restrictions of use	Academic – non-commercial use
Access medium	Downloadable
Download location	http://speechlab.tmit.bme.hu/CESAR/words_hu_v31.zip

Fee	1000 EURO	
Signatories	Henk Tamás	
	Position	Head of Department
	Contact	Magyar tudósok körútja 2. H-1117 Budapest henk@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics henk@tmit.bme.hu
Distribution rights holder	Gábor Olasz	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olasz@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olasz@tmit.bme.hu

Metadata

Creation date	2011-11-17	
Metadata creators	Mátyás Bartalis	
	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest bartalis@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bartalis@tmit.bme.hu
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2011-11-18	
Revision	2011-11-18	

Validation

--	--

Validated	True	
Type	Content	
Mode	Manual	
Details	Manually checked the sound boundaries in the corpora	
Validator	Gábor Olszy	
	Position	Professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu

Usage

Access tool	Internet browser	
Foreseen use	Human use	
	NLP applications	
NLP-specific use	Knowledge discovery	
	Linguistic research	
	Speech analysis	
Actual uses	NLP applications	
	NLP-specific use	Speech analysis
	Reports	The Phonetician 97-98. 2008/2011 http://www.isphs.org/
	Derived resource	First hungarian word level speech database
	Actual use details	This speech database contains words. The segmental level of speech (sound combinations) can be studied on acoustic level (sound spectrogram, formant structures, timing data, sound intensity)

Resource creation

Resource creator	Gábor Olszy	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics

	Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
Creation start date	2007-01-01
Creation end date	2011-08-20

Resource documentation

Reports	The Phonetician 97-98. 2008/2011 http://www.isphs.org/ http://speechlab.tmit.bme.hu/CESAR/words_hu_description_en.pdf http://speechlab.tmit.bme.hu/CESAR/words_hu_description_hu.pdf
Samples location	http://magyarbeszed.tmit.bme.hu/cvvc/index.php?hl=en

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	3936 words
Size	3936 words	
Character encoding	ISO-8859-2	
Creation	Original source	research
	Creation mode	Manual

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	3681 seconds
Audio size	3681 seconds (3681 seconds of effective speech in 3681 seconds of audio content)	
Audio content	Speech items	Isolated words
	Noise level	Low
Setting	Naturalness	Read speech
	Conversational type	Monologue
	Scenario type	Other
	Audience	No

	Interactivity	Non interactive
Audio formats	Wave/audio	
	Signal encoding	Linear PCM
	Sampling rate	22050
	Quantization	16
	Compression	False
	Number of tracks	1
	Recording quality	High
	Size	3681 seconds
Annotation	Segmentation	
	Annotated elements	Other
	Segmentation level	Phoneme
	Format	Praat TextGrid
	Tagset	http://praat.org
	Annotation mode	Mixed
	Annotation mode details	TTS for the text to sound conversion, and manual alignment of sound boundaries
	Annotation tool	Profivox development tool
	Start date	2007-01-01
	End date	2011-10-31
	Size	23541 phonemes
Recording	Device	Other
	Device details	RME Fireface800
	Platform software	soundforge_sony
	Environment	Studio
	Recorders	Gábor Olaszy
		Position professor
		Contact Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
		Organization Budapest University of Technology and Economics

			Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
Capture	Capturing device type	Studio equipment	
	Capturing device type details	AKG C-414 B-ULS	
	Person source set	Number of persons	2
		Age of persons	Adult
		Age range start	30
		Age range end	60
		Sex of persons	Mixed
		Origin of persons	Native
		Dialect accent of persons	no dialect
		Hearing impairment of persons	No
		Speaking impairment of persons	No
		Number of trained speakers	2
Creation	Original source	corpora	

2.3. Hungarian BABEL

General Information

Short name	hu-BABEL
Description	BABEL database is an EUROM1 compatible database containing records from 60 speakers distributed in 3 speakers set (many, few, very few). Paragraphs, numbers and CVC are recorded and transcribed orthographically. 10% phoneme segmentation of paragraphs (SFS format).
Identifier	203

Resource type	Corpus
URL	http://catalog.elra.info/product_info.php?products_id=577
Version	1.0
Last update	1998-12-31

Contacts

Klára Vicsi	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Availability start date	1999-01-01	

Licences

ELRA_END_USER	
Restrictions of use	Other
Access medium	CD-ROM

Metadata

Creation date	2011-11-17
Metadata creators	György Szaszák

	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2011-11-24	
Revision	2011-11-24	

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Speech analysis Speech recognition	
Actual uses	Human use	
	NLP-specific use	Speech analysis

Resource creation

Resource creator	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Creation start date	1995-01-01	
Creation end date	1998-12-31	

Resource documentation

Reports	Roach, P. - S. Arnfield, W. - Barry, J. - Baltova, M. - Boldea, A. - Fourcin, W. - Gonet, R. - Gubrynowicz, E. - Hallum, L. - Lamel, K. - Marasek, A. - Marchal, E. - Meister, E. -
----------------	---

	Vicsi, K.: BABEL: An Eastern European Multi-language database. International Conference on Speech and Language Processing, Philadelphia, 1996.
--	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	1075 utterances
Size	1075 utterances	
Character encoding	MacDingbat	
Creation	Original source	research
	Creation mode	Manual

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	1.3 hours
Audio size	1.3 hours	
Audio content	Speech items	Isolated words Natural numbers Other
	Noise level	Low
Setting	Naturality	Read speech
	Conversational type	Monologue
	Scenario type	Other
	Audience	No
	Interactivity	Non interactive
Audio formats	Wave/audio	
	Signal encoding	Linear PCM
	Sampling rate	44100
	Quantization	16
	Compression	False
	Number of	1

	tracks			
	Recording quality	High		
	Size	1.3 hours		
Annotation	Speech annotation – orthographic transcription			
	Annotated elements	Speaker noise		
	Segmentation level	Phoneme		
	Format	SAM V4.1		
	Annotation mode	Manual		
	Annotation mode details	annotation based on listening		
	Annotation tool	Self developed annotator tool		
	Start date	1996-01-01		
	End date	1998-12-31		
Recording	Device	Other		
	Device details	OROS AU21 board		
	Platform software	soundforge_sony		
	Environment	Studio		
Capture	Capturing device type	Studio equipment		
	Capturing device type details	BK microphone 4165 + BK amplifier 2636		
	Person source set	Number of persons	60	
		Age of persons	Adult	
		Age range start	25	
		Age range end	75	
		Sex of persons	Mixed	
		Origin of persons	Native	
		Dialect accent of	no dialect	

		persons	
		Hearing impairment of persons	No
		Speaking impairment of persons	No
		Number of trained speakers	0
Creation	Original source	corpora (phonetically rich + numbers + CVC)	

2.4. Hungarian Broadcast News Database

General Information

Short name	hu-broadcast
Description	The Hungarian Broadcast News (HBN) database was collected as a member of the Broadcast News Interest Group of COST278, the COST action on Speech and Language Interaction in Telecommunications in cooperation of 10 different institutions throughout Europe. The Hungarian material consists of 3h and 30minutes of recordings, transcribed and annotated (on audio level), using the conventions of NIST (National Institute of Standards and Technology, USA). The HBN is freely available for non-commercial research purposes, however it is not redistributable and the original video may never be broadcasted or played to public.
Identifier	204
Resource type	Corpus
URL	http://alpha.tmit.bme.hu/speech/
Version	1.1
Last update	2005-12-31

Contacts

Klára Vicsi	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Availability start date	2011-11-30	

Licences

MS-NC-NoReD-ND		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://alpha.tmit.bme.hu/speech/HBNC.php	
Signatories	Henk Tamás	
	Position	Head of Department
	Contact	Magyar tudósok körútja 2. H-1117 Budapest henk@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics henk@tmit.bme.hu
Distribution rights holder	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu

Metadata

Creation date	2011-11-24
----------------------	------------

Metadata creators	György Szaszák	
	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2011-11-24	
Revision	2011-11-24	

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Discourse analysis Speaker identification Speech analysis Speech to speech translation Speech understanding	
Actual uses	Human use	
	NLP-specific use	Speaker identification Speech analysis Speech recognition

Resource creation

Resource creator	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Creation start date	2004-12-01	

Creation end date	2005-12-31
--------------------------	------------

Resource documentation

Reports	Janez Zibert, France Mihelic, Jean-Pierre Martens, Hugo Meinedo, Joao Neto, Laura Docio, Carmen Garcia-Mateo, Petr David, Jan Nouza, Matus Pleva, Anton Cizmar, Andrej Zgank, Zdravko Kacic, Csaba Teleki, Klara Vicsi: The COST278 Broadcast News Segmentation and Speaker Clustering Evaluation - Overview, Methodology, Systems, Results. In: Interspeech 2005 - Eurospeech: 9th European Conference on Speech Communication and Technology. Lisboa, Portugal, 2005, ISCA, pp. 629-632.
----------------	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	25257 words
Size	25257 words, 952 turns	
Character encoding	MacDingbat	
Creation	Original source	TV broadcast transcription
	Creation mode	Manual

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	3.25 hours
Audio size	3.25 hours (194 minutes of audio content)	
Audio content	Speech items	Other
	Noise level	Low
Setting	Naturality	Spontaneous
	Conversational type	Multilogue
	Scenario type	Other
	Audience	Large public
	Interactivity	Non interactive
Audio formats	Wave/audio	

	Signal encoding	Linear PCM
	Sampling rate	16000
	Quantization	16
	Compression	False
	Number of tracks	1
	Recording quality	Very high
	Size	194 minutes
Annotation	Speech annotation – orthographic transcription	
	Annotated elements	Background noise
	Segmentation level	Utterance
	Format	Transcriber 1.4.2
	Annotation mode	Manual
	Annotation mode details	Segmentated into speaker turns, speaker noises, overlapping music are also marked.
	Annotation tool	Transcriber 1.4.2
	Start date	2005-05-01
	End date	2005-11-30
	Size	952 turns
Recording	Device	Other
	Device details	Pinnacle DV500
	Platform software	other
	Environment	Studio
Creation	Original source	TV broadcast (public service news)

2.5. Sound Gesture Database

General Information

Short name	hu-gesture
Description	Audio lexicon of sound gestures.
Identifier	205
Resource type	Lexical conceptual resource
Lexical conceptual	Lexicon

resource type	
URL	http://alpha.tmit.bme.hu/speech/
Version	1.0
Last update	2010-07-31

Contacts

Klára Vicsi	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Availability start date	2011-11-30	

Licences

MSCommons-BY-NC-SA	
Restrictions of use	Academic – non-commercial use
Access medium	Downloadable
Download location	http://alpha.tmit.bme.hu/speech/gestures_license.php

Metadata

Creation date	2011-11-17

Metadata creators	György Szaszák	
	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2011-11-24	
Revision	2011-11-24	

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Emotion recognition Talking head synthesis	
Actual uses	Human use	
	NLP-specific use	Speech analysis

Resource creation

Resource creator	Anita Czira	
	Position	MSc student
	Contact	Magyar tudósok körútja 2. H-1117 Budapest viczi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics viczi@tmit.bme.hu
Creation start date	2008-01-01	
Creation end date	2010-07-31	

Resource documentation

--	--

Reports	Vicsi Klára, Sztahó Dávid, Kiss Gábor, Czira Anita: Spontán beszédben rejlő nem verbális hangjelenségek - érzelmek, hanggesztusok - vizsgálata. In: Tanács Attila, Vincze Veronika (szerk.) MSZNY 2010: VII: Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Magyarország, pp. 249-260.
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Lexicon
---	---------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Size	100 lexical types	
Character encoding	UTF-8	

Audio recordings

Media type	audio	
Audio formats	Wave/audio	
	Signal encoding	Linear PCM
	Sampling rate	16000
	Number of tracks	1
	Recording quality	High

2.6. Hungarian Speech Emotion Database

General Information

Short name	hu-emotion
Description	Emotionally labelled speech database. Utterances are labelled according to basic emotion categories.
Identifier	206
Resource type	Corpus
URL	http://alpha.tmit.bme.hu/speech/
Version	1.0
Last update	2011-11-28

Contacts

Klára Vicsi	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Availability start date	2011-11-30	

Licences

MS-NC-NoReD-ND		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://alpha.tmit.bme.hu/speech/mtuba.php	
Signatories	Henk Tamás	
	Position	Head of Department
	Contact	Magyar tudósok körútja 2. H-1117 Budapest henk@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics

		henk@tmit.bme.hu
Distribution rights holder	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu

Metadata

Creation date	2011-11-17	
Metadata creators	György Szaszák	
	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2011-11-18	
Revision	2011-11-28	

Usage

Foreseen use	Human use	
	NLP applications	
NLP-specific use	Emotion recognition	
Actual uses	Human use	
	NLP-specific use	Speech analysis

Resource creation

Resource creator	Klára Vicsi	

	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Creation start date	2008-01-01	
Creation end date	2010-07-31	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	66400 utterances
Size	66400 utterances	
Character encoding	UTF-8	
Creation	Original source	speech
	Creation mode	Manual

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	66400 utterances
Audio size	66400 utterances (5 hours of audio content)	
Audio content	Speech items	Free speech
	Noise level	Medium
Setting	Naturality	Spontaneous
	Conversational type	Monologue
	Scenario type	Other
	Audience	Few
Audio formats	Wave/audio	

	Signal encoding	A law	
	Sampling rate	8000	
	Quantization	8	
	Compression	False	
	Number of tracks	1	
	Recording quality	Medium	
	Size	5 hours	
Annotation	Semantic annotation – emotions		
	Annotated elements	Other	
	Segmentation level	Utterance	
	Annotation mode	Manual	
	Annotation mode details	annotation based on listening	
Recording	Device	Hard disk	
	Environment	Office	
Capture	Person source set	Number of persons	1000
		Age of persons	Adult
		Age range start	20
		Age range end	100
		Sex of persons	Mixed
		Origin of persons	Native
		Dialect accent of persons	true
		Hearing impairment of persons	No
		Speaking impairment of persons	No

		Number of trained speakers	0
Creation	Original source	telephone conversations	

2.7. Hungarian MTBA

General Information

Short name	hu-MTBA
Description	Hungarian MTBA is issued from a project for the creation of the fixed line and mobil telephone voices based Hungarian speech database. The goal of the project was collecting speech telephone database, in which some major dialectal variants are represented. This database provided a realistic base both for the training and testing of the present-day teleservices, and - because of the phonetically richness - the training of real speaker independent speech recognizers. The database contains records based on the definition in SpeechDatE for the dialectal, age and sex balance and vocabulary. Important and different from the SpeechDatE database is, that the phonetically rich sentences and words have been segmented and labelled at phoneme level. Thus the database gives possibility to train phoneme based recognizers. During planning the corpus, we took into consideration not only the variety of the dialectal aspects, but the special characteristics of Hungarian language too. Since the Hungarian is an agglutinative language, we needed to create a larger vocabulary in some categories, than it was mandatory. We tried to pay an extra attention to the topic 'phonetically rich sentences and words', to create a phonetically well balanced speech database for text independent speech recognizers. A detailed statistical analysis was prepared to examine the statistics of phonemes, diphones, triphones and syllables.
Identifier	207
Resource type	Corpus
URL	http://alpha.tmit.bme.hu/speech/hdbMTBA.php
Version	1.0
Last update	2003-12-31

Contacts

Klára Vicsi	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Availability start date	2012-07-02	

Licences

MS-C-NoReD-ND-FF	
Restrictions of use	No redistribution
Access medium	CD-ROM
Fee	6500 EUR
Attribution text	In case of interest, please contact the IPR-holder specified below.

Metadata

Creation date	2012-07-02	
Metadata creators	György Szaszák	
	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2012-07-02	

Usage

Foreseen use	Human use
---------------------	-----------

	NLP applications	
NLP-specific use	Person recognition Speech analysis Speech recognition Spoken dialogue systems	
Actual uses	Human use	
	NLP-specific use	Spoken dialogue systems

Resource creation

Resource creator	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Creation start date	2001-01-01	
Creation end date	2003-12-31	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Size	5 hours	
Creation	Original source	research
	Creation mode	Manual

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	5 hours

Audio size	5 hours		
Audio content	Speech items	Isolated words Natural numbers Phonetically rich sentences	
	Noise level	Medium	
Audio formats	Wave/audio		
	Signal encoding	Linear PCM	
	Sampling rate	8000	
	Quantization	16	
	Compression	False	
	Number of tracks	1	
	Recording quality	High	
	Size	5 hours	
Annotation	Segmentation		
	Annotated elements	Speaker noise	
	Segmentation level	Phoneme	
	Annotation mode	Manual	
	Annotation mode details	annotation based on listening	
	Annotation tool	Self developed annotator tool	
	Start date	2002-01-01	
	End date	2003-12-31	
Capture	Person source set	Number of persons	500
		Age of persons	Adult
		Age range start	18
		Age range end	99
		Sex of persons	Mixed
		Origin of persons	Native
		Dialect	varied, balanced

		accent of persons	
		Hearing impairment of persons	No
		Speaking impairment of persons	No
		Number of trained speakers	0
Creation	Original source	corpora (phonetically rich + numbers)	

2.8. Hungarian MRBA

General Information

Short name	hu-MRBA
Description	<p>The Hungarian Reference Speech Database (MRBA) was developed at the Laboratory of Speech Acoustics of the Budapest University of Technology and Economics (BME) in collaboration with the Institute of Informatics of the University of Szeged [1]. The main goal was to develop a speech database that contains continuous read speech, so that the database can be used for training and testing of PC-based automatic speech recognisers. During the planning of the corpus, we took into consideration the special characteristics of Hungarian language. Since the Hungarian is an agglutinative language, we needed to create a larger vocabulary in some categories, than it is mandatory. We tried to pay an extra attention to the topic 'phonetically rich sentences and words', to create a phonetically well balanced speech database for text independent speech recognizers. A detailed statistical analysis was prepared to examine the statistics of phonemes, diphones, triphones and syllables. In this way every speaker had to read 12 different sentences and 12 different words, that had no connection with the sentences. The database contains utterances read by 332 different speakers. The utterances were recorded in acoustically different locations, such as office, laboratories, home. The database contains utterances recorded simultaneously with two different systems. One of these systems was considered the reference system. This reference system contained a laptop, an external sound card and a good quality condenser microphone. The reference system was unchanged until the database was finished. In case of the other system, we changed the microphones, sound cards, PC-s. To cover the dialects spoken in Hungary, we made records in four different locations of the country and we took into consideration the gender and age of speakers, so the database has balanced distribution over gender, age and dialects. Every spoken utterance has been labeled, so every wave (16kHz, 16bit, mono) file has a label file, which contains informations about the parameters of the record and the orthographical transcription of the spoken material. Almost one third of the database (100 speakers' utterances) was manually segmented and labelled at phoneme level, using SAMPA codes.</p>
Identifier	208
Resource type	Corpus
URL	http://alpha.tmit.bme.hu/speech/hdbMRBA.php
Version	1.0

Last update	2007-12-31
--------------------	------------

Contacts

Klára Vicsi	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Availability start date	2012-07-02	

Licences

MS-C-NoReD-ND-FF	
Restrictions of use	No redistribution
Access medium	CD-ROM
Fee	6500 EUR
Attribution text	In case of interest, please contact the IPR-holder specified below.

Metadata

Creation date	2012-07-02	
Metadata creators	György Szaszák	
	Position	research fellow

	Contact	Magyar tudósok körútja 2. H-1117 Budapest szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2012-07-02	

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Person recognition Speech analysis Speech recognition Spoken dialogue systems	
Actual uses	Human use	
	NLP-specific use	Speech recognition

Resource creation

Resource creator	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Creation start date	2004-01-01	
Creation end date	2007-12-31	

Texts

Media type	text
Linguality type	Monolingual

Languages	Hungarian	
	Language ID	HU
Size	6 hours	
Creation	Original source	research
	Creation mode	Manual

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	6 hours
Audio size	6 hours	
Audio content	Speech items	Isolated words Phonetically rich sentences
	Noise level	Medium
Audio formats	Wave/audio	
	Signal encoding	Linear PCM
	Sampling rate	16000
	Quantization	16
	Compression	False
	Number of tracks	1
	Recording quality	High
	Size	6 hours
Annotation	Segmentation	
	Annotated elements	Speaker noise
	Segmentation level	Phoneme
	Annotation mode	Manual
	Annotation mode details	annotation based on listening
	Annotation tool	Self developed annotator tool
	Start date	2005-01-01

	End date	2007-06-30	
Capture	Person source set	Number of persons	332
		Age of persons	Adult
		Age range start	18
		Age range end	99
		Sex of persons	Mixed
		Origin of persons	Native
		Dialect accent of persons	varied, balanced
		Hearing impairment of persons	No
		Speaking impairment of persons	No
		Number of trained speakers	0
Creation	Original source	corpora (phonetically rich and balanced)	

2.9. Hungarian Phone Speech Call Center Database

General Information

Short name	hu-MTUBA
Description	The Hungarian Phone Speech Call Center Database is a telephone speech database containing discourses between the operators of a service provider company and its clients. Orthographic transcription is provided. Emotions are also labelled. A derivative of this database called Hungarian Speech Emotion Database is also available from META-SHARE, with free academic use.
Identifier	209
Resource type	Corpus
URL	http://alpha.tmit.bme.hu/speech/mtuba.php
Version	2.0
Last update	2012-02-29

Contacts

Klára Vicsi	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Availability start date	2012-07-02	

Licences

MS-C-NoReD-ND-FF	
Restrictions of use	No derivatives No redistribution
Access medium	CD-ROM
Fee	12000 EUR
Attribution text	A derivative of this database called Hungarian Speech Emotion Database is also available from META-SHARE, with free academic use, however, without orthographic transcription.
MS-NC-NoReD-ND-FF	
Restrictions of use	No derivatives No redistribution
Access medium	CD-ROM
Fee	8000 EUR

Attribution text	A derivative of this database called Hungarian Speech Emotion Database is also available from META-SHARE, with free academic use, however, without orthographic transcription.
-------------------------	--

Metadata

Creation date	2012-07-02
Metadata creators	György Szaszák
	Position research fellow
	Contact Magyar tudósok körútja 2. H-1117 Budapest szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu
Source	CESAR
Metadata language ID	en-us
Metadata last date updated	2012-07-02

Usage

Foreseen use	Human use NLP applications
NLP-specific use	Emotion recognition Person recognition Speech analysis Speech recognition Spoken dialogue systems
Actual uses	Human use
	NLP-specific use Emotion recognition Speech recognition

Resource creation

Resource creator	Klára Vicsi
	Position professor
	Contact Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/

	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Creation start date	2009-01-01	
Creation end date	2012-02-29	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Size	1038 utterances	
Creation	Original source	research
	Creation mode	Manual

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	1038 utterances
Audio size	1038 utterances 3.5 hours 1.8 gb	
Audio content	Noise level	Medium
Audio formats	Wave/audio	
	Signal encoding	A law
	Sampling rate	8000
	Quantization	8
	Compression	False
	Number of tracks	1
	Recording quality	Medium
	Size	3.5 hours
Annotation	Semantic annotation – emotions	
	Annotated elements	Other

	Segmentation level	Phrase	
	Annotation mode	Manual	
	Annotation mode details	annotation based on listening	
	Annotation tool	Praat	
	Start date	2009-01-01	
	End date	2012-02-29	
Capture	Person source set	Number of persons	1038
		Age of persons	Adult
		Age range start	18
		Age range end	99
		Sex of persons	Mixed
		Origin of persons	Native
		Dialect accent of persons	varied
		Hearing impairment of persons	No
		Speaking impairment of persons	No
		Number of trained speakers	0

2.10. Hungarian BABEL phonetic segmentation and syntactic and prosodic analysis

General Information

Short name	hu-BABEL-addons
Description	BABEL database is an EUROM1 compatible database containing records from 60 speakers distributed in 3 speakers set (many, few, very few). The resource is available under META-SHARE via ELRA, this supplement is an add-on to the database. In order to

	use it, the BABEL database is necessary.
Identifier	210
Resource type	Corpus
URL	http://catalog.elra.info/product_info.php?products_id=577
Version	1.0
Last update	2012-03-31

Contacts

Klára Vicsi	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Availability start date	2012-08-01	

Licences

MS-NC-NoReD	
Restrictions of use	No redistribution
Access medium	Downloadable
Download location	http://alpha.tmit.bme.hu/speech/hdbbabel.php
Attribution text	BABEL database itself is necessary, available via ELRA-END-USER licence.

Metadata

Creation date	2012-07-02	
Metadata creators	György Szaszák	
	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2012-07-02	

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Speech analysis Speech understanding	
Actual uses	Human use	
	NLP-specific use	Speech understanding

Resource creation

Resource creator	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Creation start date	2007-01-01	
Creation end date	2012-03-31	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	330 utterances
Size	330 utterances	
Character encoding	UTF-8	
Creation	Original source	research
	Creation mode	Manual

2.11. Di-phone database for text-to-speech conversion

General Information

Short name	Di-phone-hu
Description	The Di-phone set (labelled wave form items) for Hungarian contains combinations of 38 sounds for TTS conversion. Besides the Di-phone set can be used for educational purposes and in speech research.
Identifier	211
Resource type	Corpus
Version	1.0
Revision	Waves, texts, sound boundaries
Last update	2012-07-09

Contacts

Géza Németh	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest nemeth@tmit.bme.hu http://speechlab.tmit.bme.hu
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu

Distribution

Availability	Available – restricted use
---------------------	----------------------------

IPR holder	Gábor Olszy	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
Availability start date	2012-07-15	

Licences

CLARIN_RES		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://speechlab.tmit.bme.hu/CESAR/diphone_hu.zip	
Fee	1000 EURO	
Signatories	Henk Tamás	
	Position	Head of Department
	Contact	Magyar tudósok körútja 2. H-1117 Budapest henk@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics henk@tmit.bme.hu
Distribution rights holder	Gábor Olszy	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu

Metadata

Creation date	2012-07-09

Metadata creators	Mátyás Bartalis	
	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest bartalis@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bartalis@tmit.bme.hu
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2012-07-09	
Revision	2012-07-09	

Validation

Validated	True	
Type	Content	
Mode	Manual	
Details	Manually checked the sound boundaries in the corpora	
Validator	Bálint Tóth	
	Position	Ph.D. candidate
	Contact	Magyar tudósok körútja 2. H-1117 Budapest toth.b@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics toth.b@tmit.bme.hu

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Speech analysis Speech synthesis Talking head synthesis	
Actual uses	NLP applications	
	NLP-specific	Speech synthesis

	use	Talking head synthesis
	Reports	International Journal of Speech Technology, Kluwer, 2000
	Derived resource	Hungarian di-phone speech synthesizer
	Actual use details	The Profivox hungarian di-phone TTS uses a database based on this resource

Resource creation

Resource creator	Gábor Olasz	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
Creation start date	2002-01-01	
Creation end date	2012-06-30	

Resource documentation

Reports	Olasz G. - Gépi beszédeltés információs rendszerekhez Magyarországon. AKUSZTIKAI SZEMLE III:(1-3) pp. 4-13. (1999) http://speechlab.tmit.bme.hu/CESAR/diphone_hu_description_en.pdf http://speechlab.tmit.bme.hu/CESAR/diphone_hu_description_hu.pdf http://speechlab.tmit.bme.hu/CESAR/diphone_hu_script.pdf
Samples location	http://speechlab.tmit.bme.hu/profivox-tts-demo/MAGYAR_SZOVEGFELOLVASO/noi_hangok/Veronika/

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	1455 words
Size	1455 words	
Character encoding	ISO-8859-2	
Creation	Original source	research
	Creation mode	Manual

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	1646 seconds
Audio size	1646 seconds (1646 seconds of effective speech in 1646 seconds of audio content)	
Audio content	Speech items	Isolated words
	Noise level	Low
Setting	Naturality	Read speech
	Conversational type	Monologue
	Scenario type	Other
	Audience	No
	Interactivity	Non interactive
Audio formats	Wave/audio	
	Signal encoding	Linear PCM
	Sampling rate	44100
	Quantization	16
	Compression	False
	Number of tracks	1
	Recording quality	High
	Size	1646 seconds
Annotation	Segmentation	
	Annotated elements	Other
	Segmentation level	Phoneme
	Format	Praat TextGrid
	Tagset	http://praat.org
	Annotation mode	Mixed
	Annotation mode details	TTS for the text to sound conversion, and manual alignment of sound boundaries
	Annotation tool	Profivox development tool

Recording	Start date	2007-01-01	
	End date	2012-06-30	
	Size	10693 phonemes	
	Device	Other	
	Device details	RME Fireface800	
Capture	Platform software	soundforge_sony	
	Environment	Studio	
	Recorders	Gábor Olaszy	
		Position	professor
		Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
		Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
	Capturing device type	Studio equipment	
	Capturing device type details	AKG C-414 B-ULS	
	Person source set	Number of persons	1
		Age of persons	Adult
		Age range start	24
		Age range end	24
		Sex of persons	Female
		Origin of persons	Native
		Dialect accent of persons	no dialect
		Hearing impairment of persons	No

		Speaking impairment of persons	No
		Number of trained speakers	0
Creation	Original source	corpora	

2.12. Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection

General Information

Short name	ParallelSpeech-hu
Description	Phonetically balanced sentence set read by 10 speakers.
Identifier	212
Resource type	Corpus
Version	1.0
Revision	Waves, texts, sound boundaries
Last update	2012-07-09

Contacts

Géza Németh	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest nemeth@tmit.bme.hu http://speechlab.tmit.bme.hu
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics nemeth@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Gábor Olaszy	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu

	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
Availability start date	2012-07-15	

Licences

CLARIN_RES		
Restrictions of use	Academic – non-commercial use	
Access medium	DVD-R	
Fee	4000 EURO pro speaker	
Signatories	Henk Tamás	
	Position	Head of Department
	Contact	Magyar tudósok körútja 2. H-1117 Budapest henk@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics henk@tmit.bme.hu
Distribution rights holder	Géza Németh	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest nemeth@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics nemeth@tmit.bme.hu

Metadata

Creation date	2012-07-09	
Metadata creators	Mátyás Bartalis	
	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest bartalis@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics

	Department of Telecommunications and Media Informatics bartalis@tmit.bme.hu
Source	CESAR
Metadata language ID	en-us
Metadata last date updated	2012-07-09
Revision	2012-07-09

Validation

Validated	True	
Type	Content	
Mode	Manual	
Details	Manually checked annotation and labeling of the sound and word boundaries in the corpora	
Validator	Tamás Gábor Csapó	
	Position	Ph.D. candidate
	Contact	Magyar tudósok körútja 2. H-1117 Budapest csapot@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics csapot@tmit.bme.hu

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Knowledge discovery Linguistic research Speech analysis Speech synthesis	
Actual uses	NLP applications	
	NLP-specific use	Speech analysis Speech synthesis
	Derived resource	First hungarian precisely labelled parallel speech database collection
	Actual use details	This speech database contains 2000 sentences. Each speaker read this sentence set. This parallel speech database is used to train HMM based TTS and for unit selection TTS.

Resource creation

Resource creator	Csaba Zainkó	
	Position	Ph.D. lecturer
	Contact	Magyar tudósok körútja 2. H-1117 Budapest zainko@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics zainko@tmit.bme.hu
	Tamás Böhm	
	Position	Ph.D. lecturer
	Contact	Magyar tudósok körútja 2. H-1117 Budapest bohm@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bohm@tmit.bme.hu
Creation start date	2009-07-01	
Creation end date	2012-09-30	

Resource documentation

Reports	http://speechlab.tmit.bme.hu/CESAR/ParallelSpeech_hu_description_en.pdf http://speechlab.tmit.bme.hu/CESAR/ParallelSpeech_hu_description_hu.pdf
----------------	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	19658 sentences
Size	19658 sentences	
Character encoding	ISO-8859-2	
Creation	Original source	research
	Creation mode	Manual

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	91924 seconds
Audio size	91924 seconds (91924 seconds of effective speech in 91924 seconds of audio content)	
Audio content	Speech items	Phonetically balanced sentences
	Noise level	Low
Setting	Naturality	Read speech
	Conversational type	Monologue
	Scenario type	Other
	Audience	No
	Interactivity	Non interactive
Audio formats	Wave/audio	
	Signal encoding	Linear PCM
	Sampling rate	44100
	Quantization	16
	Compression	False
	Number of tracks	1
	Recording quality	High
	Size	91924 seconds
Annotation	Segmentation	
	Annotated elements	Other
	Segmentation level	Phoneme
	Format	Praat TextGrid
	Tagset	http://praat.org
	Annotation mode	Mixed
	Annotation mode details	TTS for the text to sound conversion, and manual alignment of sound boundaries
	Annotation tool	Profivox development tool

	Start date	2009-08-01	
	End date	2012-09-30	
	Size	831941 phonemes	
	Annotators	Klára Laczkó	
		Position	staff member
		Contact	Magyar tudósok körútja 2. H-1117 Budapest klara@tmit.bme.hu http://www.tmit.bme.hu
		Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics klara@tmit.bme.hu
Recording	Device	Other	
	Device details	RME Fireface800	
	Platform software	soundforge_sony	
	Environment	Studio	
	Recorders	Mátyás Bartalis	
		Position	research fellow
		Contact	Magyar tudósok körútja 2. H-1117 Budapest bartalis@tmit.bme.hu http://speechlab.tmit.bme.hu
		Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bartalis@tmit.bme.hu
Capture	Capturing device type	Studio equipment	
	Capturing device type details	AKG C-414 B-ULS	
	Person source set	Number of persons	10
		Age of persons	Adult
		Age range start	26
		Age range end	60

		Sex of persons	Mixed
		Origin of persons	Native
		Dialect accent of persons	no dialect
		Hearing impairment of persons	No
		Speaking impairment of persons	No
		Number of trained speakers	2
Creation	Original source	corpora	

2.13. Read speech database in Hungarian

General Information

Short name	ReadSpeech-hu
Description	The read speech database contains sentences from weather forecast news. The sentence collection represents the four seasons. This database can be used for analysing speech characteristics in weather forecast news and also as the basic speech database of a corpus based Concept-to-Speech system.
Identifier	213
Resource type	Corpus
Version	1.0
Revision	Waves, texts, sound boundaries
Last update	2012-07-09

Contacts

Géza Németh	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest nemeth@tmit.bme.hu http://speechlab.tmit.bme.hu
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics

	nemeth@tmit.bme.hu
--	--

Distribution

Availability	Available – restricted use	
IPR holder	Gábor Olszy	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
Availability start date	2012-07-15	

Licences

CLARIN_RES		
Restrictions of use	Academic – non-commercial use	
Access medium	DVD-R	
Fee	40000 EURO	
Signatories	Henk Tamás	
	Position	Head of Department
	Contact	Magyar tudósok körútja 2. H-1117 Budapest henk@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics henk@tmit.bme.hu
Distribution rights holder	Géza Németh	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest nemeth@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics nemeth@tmit.bme.hu

Metadata

Creation date	2012-07-09	
Metadata creators	Mátyás Bartalis	
	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest bartalis@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bartalis@tmit.bme.hu
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2012-07-09	
Revision	2012-07-09	

Validation

Validated	True	
Type	Content	
Mode	Manual	
Details	Manually checked the sound and the text sentence by sentence	
Validator	Tamás Gábor Csapó	
	Position	Ph.D. candidate
	Contact	Magyar tudósok körútja 2. H-1117 Budapest csapot@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics csapot@tmit.bme.hu
	Bálint Pál Tóth	
	Position	Ph.D. candidate
	Contact	Magyar tudósok körútja 2. H-1117 Budapest toth.b@tmit.bme.hu

	http://www.tmit.bme.hu
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics toth.b@tmit.bme.hu
Tamás Böhm	
Position	Ph.D. lecturer
Contact	Magyar tudósok körútja 2. H-1117 Budapest bohm@tmit.bme.hu http://www.tmit.bme.hu
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bohm@tmit.bme.hu

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Knowledge discovery Linguistic research Speech analysis Speech synthesis	
Actual uses	NLP applications	
	NLP-specific use	Speech analysis Speech synthesis
	Derived resource	Large corpus focused on the weather forecast
	Actual use details	The first automatic TTS based Hungarian weather forecast application (www.metnet.hu) based on this database

Resource creation

Resource creator	Csaba Zainkó	
	Position	Ph.D. lecturer
	Contact	Magyar tudósok körútja 2. H-1117 Budapest zainko@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics zainko@tmit.bme.hu
	Tamás Böhm	

	Position	Ph.D. lecturer
	Contact	Magyar tudósok körútja 2. H-1117 Budapest bohm@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bohm@tmit.bme.hu
Creation start date	2005-01-01	
Creation end date	2012-07-10	

Resource documentation

Reports	http://www.springerlink.com/content/mr6m71133887823m http://speechlab.tmit.bme.hu/CESAR/ReadSpeech_hu_description_en.pdf
----------------	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	5529 sentences
Size	5529 sentences	
Character encoding	ISO-8859-2	
Creation	Original source	research
	Creation mode	Manual

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	36822 seconds
Audio size	36822 seconds (36822 seconds of effective speech in 36822 seconds of audio content)	
Audio content	Speech items	Other
	Noise level	Low
Setting	Naturality	Read speech
	Conversational	Monologue

	type	
	Scenario type	Other
	Audience	No
	Interactivity	Non interactive
Audio formats	Wave/audio	
	Signal encoding	Linear PCM
	Sampling rate	44100
	Quantization	16
	Compression	False
	Number of tracks	1
	Recording quality	High
	Size	36822 seconds
Annotation	Segmentation	
	Annotated elements	Other
	Segmentation level	Phoneme
	Format	Praat TextGrid
	Tagset	http://praat.org
	Annotation mode	Mixed
	Annotation mode details	TTS for the text to sound conversion, and manual alignment of sound boundaries
	Annotation tool	Profivox development tool
	Start date	2005-01-01
	End date	2012-07-10
	Size	466112 phonemes
	Annotators	Mátyás Bartalis
		Position research fellow
		Contact Magyar tudósok körútja 2. H-1117 Budapest bartalis@tmit.bme.hu http://speechlab.tmit.bme.hu
		Organization Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bartalis@tmit.bme.hu

		Klára Laczkó	
		Position	staff member
		Contact	Magyar tudósok körútja 2. H-1117 Budapest klara@tmit.bme.hu http://www.tmit.bme.hu
		Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics klara@tmit.bme.hu
Recording	Device	Other	
	Device details	RME Fireface800	
	Platform software	soundforge_sony	
	Environment	Studio	
	Recorders	Mátyás Bartalis	
		Position	research fellow
		Contact	Magyar tudósok körútja 2. H-1117 Budapest bartalis@tmit.bme.hu http://speechlab.tmit.bme.hu
		Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bartalis@tmit.bme.hu
Capture	Capturing device type	Studio equipment	
	Capturing device type details	AKG C-414 B-ULS	
	Person source set	Number of persons	1
		Age of persons	Adult
		Sex of persons	Female
		Origin of persons	Native
		Dialect accent of persons	no dialect
		Hearing	No

		impairment of persons	
		Speaking impairment of persons	No
		Number of trained speakers	1
Creation	Original source	corpora	

2.14. Hungarian Book (Egri csillagok/Eclipse of the Crescent Moon by Géza Gárdonyi) Reading Speech and Aligned Text Selection Database

General Information

Description	Database of portions of text and audio version of a Hungarian novel. (The audio data is not stored in this database, but can be freely downloaded from librivox.org .) The recordings are segmented between speech pauses, which not necessarily correspond to sentence boundaries. The reading is mostly, but not completely accurate. Hence, an automatic speech recognizer was utilized to choose only those segments, where there is a high match between the automatic recognition result and the original text. Thus the database comprises only those segments that are considered to have a reliable transcription. The database can be applied in speech technology research, phonetic, phonological research and for developing and testing speech recognition systems.
Identifier	214
Resource type	Corpus
Version	1.1

Contacts

Péter Mihajlik	
Contact	Magyar tudósok körútja 2. H-1117 Budapest mihajlik@tmit.bme.hu http://alpha.tmit.bme.hu/~mihajlik/
Organization	THINKTech Research Center non-profit LLC info@thinktech.hu

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

--

CC-BY		
Access medium	Downloadable	
Download location	http://thinktech.hu/public/lang-res/HunAuB-ECM-text/	
Attribution text	Éva Székely, Tamás Gábor Csapó, Bálint Tóth, Péter Mihajlik, Julie Carson-Berndsen: Synthesizing Expressive Speech from Amateur Audiobook Recordings. In: IEEE Workshop on Spoken Language Technology. Miami, United States of America, 02/12/2012-05/12/2012. Miami: Paper 1.	
Signatories	Péter Mihajlik	
	Contact	Magyar tudósok körútja 2. H-1117 Budapest mihajlik@tmit.bme.hu http://alpha.tmit.bme.hu/~mihajlik/
	Organization	THINKTech Research Center non-profit LLC info@thinktech.hu

Metadata

Creation date	2012-07-12	
Metadata creators	András Balog	
	Position	engineer
	Contact	abalog@aitia.ai
	Organization	THINKTech Research Center non-profit LLC info@thinktech.hu
Source	METANET4U	
Metadata language ID	en-us	
Metadata last date updated	2013-01-23	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	34438 words
Size	237000 phonemes	
Character encoding	UTF-8	
Creation	Original source	http://mek.oszk.hu/00600/00656/index.phtml
	Creation mode	Automatic

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Audio size	617 mb (5 hours of audio content)	
Audio content	Speech items	Free speech
	Noise level	Low
Setting	Naturality	Read speech
	Conversational type	Monologue
	Scenario type	Other
	Audience	No
	Interactivity	Non interactive
Audio formats	Audio/mp3	
	Signal encoding	Other
	Sampling rate	44100
	Quantization	16
	Compression	True
	Compression name	Mp3
	Compression loss	True
	Number of tracks	1
	Recording quality	High
Annotation	Alignment	
	Segmentation level	Other
	Format	plain txt
	Annotation mode	Mixed
	Annotation mode details	The original text of the book was downloaded from the web page of the Hungarian Electronic Library. The reading of the text is mostly, but not completely correct. Hence, recordings were transcribed and segmented by an automatic speech recognizer, and were compared with the original text. The given database comprises only those segments where the match between automatic recognition result and original text is

		100%. The audio data is downloadable from the librivox.org . The size data above applies to the selected part of the audio book.
	Annotation tool	voXserver ASR engine, other self developed processing tools
Creation	Original source	http://librivox.org/egri-csillagok-by-geza-gardonyi/

2.15. Hungarian Poem (János vitéz/John the Valiant by Sándor Petőfi) Reading Speech and Aligned Text Selection Database

General Information

Description	Database of portions of text and audio version of a Hungarian piece of poetry. (The audio data is not stored in this database, but can be freely downloaded from librivox.org .) The recordings are segmented between speech pauses, which not necessarily correspond to sentence boundaries. The reading is mostly, but not completely accurate. Hence, an automatic speech recognizer was utilized to choose only those segments, where there is a high match between the automatic recognition result and the original text. Thus the database comprises only those segments that are considered to have a reliable transcription. The database can be applied in speech technology research, phonetic, phonological research and for developing and testing speech recognition systems.
Identifier	215
Resource type	Corpus
Version	1.1

Contacts

Péter Mihajlik	
Contact	Magyar tudósok körútja 2. H-1117 Budapest mihajlik@tmit.bme.hu http://alpha.tmit.bme.hu/~mihajlik/
Organization	THINKTech Research Center non-profit LLC info@thinktech.hu

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

CC-BY	
Access medium	Downloadable
Download location	http://thinktech.hu/public/lang-res/HunReP-SJ-text/
Attribution text	Éva Székely, Tamás Gábor Csapó, Bálint Tóth, Péter Mihajlik, Julie Carson-Berndsen: Synthesizing Expressive Speech from Amateur Audiobook Recordings. In: IEEE

	Workshop on Spoken Language Technology. Miami, United States of America, 02/12/2012-05/12/2012. Miami: Paper 1.	
Signatories	Péter Mihajlik	
	Contact	Magyar tudósok körútja 2. H-1117 Budapest mihajlik@tmit.bme.hu http://alpha.tmit.bme.hu/~mihajlik/
	Organization	THINK Tech Research Center non-profit LLC info@thinktech.hu

Metadata

Creation date	2012-07-12	
Metadata creators	András Balog	
	Position	engineer
	Contact	abalog@aitia.ai
	Organization	THINK Tech Research Center non-profit LLC info@thinktech.hu
Source	METANET4U	
Metadata language ID	en-us	
Metadata last date updated	2013-01-23	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	4436 words
Size	31000 phonemes	
Character encoding	UTF-8	
Creation	Original source	http://mek.oszk.hu/01000/01010/index.phtml
	Creation mode	Automatic

Audio recordings

Media type	audio
Linguality type	Monolingual
Languages	Hungarian

	Language ID	HU
Audio size	84 mb (44 minutes of audio content)	
Audio content	Speech items	Free speech
	Noise level	Low
Setting	Naturality	Read speech
	Conversational type	Monologue
	Scenario type	Other
	Audience	No
	Interactivity	Non interactive
Audio formats	Audio/mp3	
	Signal encoding	Other
	Sampling rate	44100
	Quantization	16
	Compression	True
	Compression name	Mp3
	Compression loss	True
	Number of tracks	1
	Recording quality	High
Annotation	Alignment	
	Segmentation level	Other
	Format	plain txt
	Annotation mode	Mixed
	Annotation mode details	The original text of the book was downloaded from the web page of the Hungarian Electronic Library. The reading of the text is mostly, but not completely correct. Hence, recordings were transcribed and segmented by an automatic speech recognizer, and were compared with the original text. The given database comprises only those segments where the match between automatic recognition result and original text is 100%. The audio data is downloadable from the librivox.org . The size data above applies to the selected part of the audio book
	Annotation tool	voXserver ASR engine, other self developed processing tools
Creation	Original source	http://librivox.org/janos-vitez-by-sandor-petofi/

2.16. Hungarian Parliamentary Speech and Aligned Text Selection Database

General Information

Description	Database of recordings and official transcripts of Hungarian parliamentary speeches. The recordings are segmented between speech pauses, which not necessarily correspond to sentence boundaries. The official transcripts are not completely accurate, since the parliamentary transcribers correct most of grammatical mistakes and speech disfluencies. Hence, an automatic speech recognizer was utilized to choose only those segments, where there is a high match between the automatic and manual transcriptions. Thus the database comprises only those segments that are considered to have a reliable transcription. The database can be applied in speech technology research, phonetic, phonological research and for developing and testing speech and speaker recognition systems.
Identifier	216
Resource type	Corpus
Version	2.0

Contacts

Péter Mihajlik	
Position	research fellow II
Contact	Magyar tudósok körútja 2. H-1117 Budapest mihajlik@tmit.bme.hu http://alpha.tmit.bme.hu/~mihajlik/
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics mihajlik@tmit.bme.hu

Distribution

Availability	Available – unrestricted use	
IPR holder	Budapest University of Technology and Economics	
	Short name	BME
	Department name	Department of Telecommunications and Media Informatics
	Contact	mihajlik@tmit.bme.hu http://www.tmit.bme.hu

Licences

CC-BY	
Access medium	CD-ROM

Signatories	Henk Tamás	
	Position	Head of Department
	Contact	Magyar tudósok körútja 2. H-1117 Budapest henk@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics henk@tmit.bme.hu

Metadata

Creation date	2012-07-10	
Metadata creators	Gellért Sárosi	
	Position	engineer
	Contact	Magyar tudósok körútja 2. H-1117 Budapest sarosi@tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics sarosi@tmit.bme.hu
Source	METANET4U	
Metadata language ID	en-us	
Metadata last date updated	2013-01-23	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Size	134 mb	
Character encoding	UTF-8	
Creation	Original source	http://www.parlament.hu/internet/plsql/ogy_naplo.naplo_ujnapok_ckl
	Creation mode	Automatic

Audio recordings

--	--

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Audio size	204 gb (1898 hours of audio content)	
Audio content	Speech items	Free speech
	Noise level	Low
Setting	Naturality	Planned
	Conversational type	Monologue
	Scenario type	Other
	Audience	Some
	Interactivity	Non interactive
Audio formats	Video/mpeg	
	Signal encoding	Other
	Quantization	16
	Compression	False
	Compression name	Mpeg
	Compression loss	True
	Number of tracks	1
	Recording quality	High
Annotation	Speech annotation – orthographic transcription	
	Segmentation level	Other
	Format	plain txt
	Annotation mode	Mixed
	Annotation mode details	The official transcripts were downloaded from the web page of the Hungarian parliament. These transcripts are not completely accurate, since the parliamentary transcribers correct most of grammatical mistakes and speech disfluencies. Hence, recordings were transcribed and segmented by an automatic speech recognizer, and were compared with the downloaded transcripts. The given text corpus comprises only those segments where the match (letter-based accuracy) between automatic and manual transcriptions is over 98%. The audio files are not part of this shared corpus. They can freely be downloaded from the Hungarian Parliament website. The size data above applies to the

		selected part of the speeches.
	Annotation tool	voXserver ASR engine, other self developed processing tools
Creation	Original source	http://www.parlament.hu/internet/plsql/ogy_naplo.naplo_ujnapok_ckl

2.17. Named entity lexical database

General Information

Short name	Name-hu
Description	This database contains the written and the spoken form of the items
Identifier	217
Resource type	Corpus
Version	1.0
Revision	Waves and texts
Last update	2013-01-22

Contacts

Géza Németh		
Position	professor	
Contact	Magyar tudósok körútja 2. H-1117 Budapest nemeth@tmit.bme.hu http://speechlab.tmit.bme.hu	
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu	

Distribution

Availability	Available – restricted use	
IPR holder	Gábor Olszy	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
Availability start	2013-01-24	

date	
------	--

Licences

CLARIN_RES		
Restrictions of use	Academic – non-commercial use	
Access medium	CD-ROM	
Fee	4000 EURO	
Signatories	Henk Tamás	
	Position	Head of Department
	Contact	Magyar tudósok körútja 2. H-1117 Budapest henk@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics henk@tmit.bme.hu
Distribution rights holder	Gábor Olaszy	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu

Metadata

Creation date	2013-01-22	
Metadata creators	Mátyás Bartalis	
	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest bartalis@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bartalis@tmit.bme.hu
Source	METANET4U	
Metadata language	en-us	

ID	
Metadata last date updated	2013-01-22
Revision	2013-01-22

Validation

Validated	True	
Type	Content	
Mode	Manual	
Details	Manually checked the text files against spelling	
Validator	Bálint Tóth	
	Position	Ph.D. candidate
	Contact	Magyar tudósok körútja 2. H-1117 Budapest toth.b@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics toth.b@tmit.bme.hu

Usage

Foreseen use	Human use	
NLP-specific use	Knowledge discovery	
Foreseen use	NLP applications	
NLP-specific use	Speech synthesis Spoken dialogue systems	
Actual uses	NLP applications	
	NLP-specific use	Speech synthesis Spoken dialogue systems
	Reports	IEEE NLP-KE 2003, pp. 238-243. (ISBN: 0-7803-7902-0)

Resource creation

Resource creator	Gábor Olaszy	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu

	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
Creation start date	2000-01-01	
Creation end date	2013-01-20	

Resource documentation

Reports	IEEE NLP-KE 2003, pp. 238-243. (ISBN: 0-7803-7902-0), http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01275906 http://speechlab.tmit.bme.hu/CESAR/name_hu_description_en.pdf http://speechlab.tmit.bme.hu/CESAR/name_hu_description_hu.pdf
----------------	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	131601 words
Size	131601 words	
Character encoding	ISO-8859-2	
Creation	Original source	research
	Creation mode	Manual

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	1717 seconds
Audio size	1717 seconds (1717 seconds of effective speech in 1717 seconds of audio content)	
Audio content	Speech items	Isolated words
	Noise level	Low
Setting	Naturality	Read speech
	Conversational type	Monologue
	Scenario type	Other
	Audience	No

	Interactivity	Non interactive	
Audio formats	Wave/audio		
	Signal encoding	Linear PCM	
	Sampling rate	22050	
	Quantization	16	
	Compression	False	
	Number of tracks	1	
	Recording quality	High	
	Size	1717 seconds	
Recording	Device	Other	
	Device details	RME Fireface800	
	Platform software	soundforge_sony	
	Environment	Studio	
	Recorders	Gábor Olaszy	
		Position	professor
		Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
		Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
Capture	Capturing device type	Studio equipment	
	Capturing device type details	AKG C-414 B-ULS	
	Person source set	Number of persons	1
		Age of persons	Adult
		Age range start	60
		Age range end	60
		Sex of	Male

		persons	
		Origin of persons	Native
		Dialect accent of persons	no dialect
		Hearing impairment of persons	No
		Speaking impairment of persons	No
		Number of trained speakers	0
Creation	Original source	corpora	

2.18. Formant database from spoken words

General Information

Short name	Hun-Formant
Description	The formant values of vowels in spoken words are defined and organised in a database. The searching facilities help the user to study the formant movements in the vowels in the function of the adjacent sounds.
Identifier	218
Resource type	Corpus
URL	http://beszedmuhely.tmit.bme.hu/formant
Version	1.0
Revision	Waves, texts, sound boundaries and three formant values of each vowels
Last update	2013-01-22

Contacts

Géza Németh	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest nemeth@tmit.bme.hu http://speechlab.tmit.bme.hu
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics nemeth@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Gábor Olszy	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
Availability start date	2013-01-24	

Licences

CLARIN_RES		
Restrictions of use	Academic – non-commercial use	
Access medium	CD-ROM	
Fee	3000 EURO	
Signatories	Henk Tamás	
	Position	Head of Department
	Contact	Magyar tudósok körútja 2. H-1117 Budapest henk@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics henk@tmit.bme.hu
Distribution rights holder	Géza Németh	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest nemeth@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics nemeth@tmit.bme.hu

Metadata

Creation date	2013-01-22	
Metadata creators	Mátyás Bartalis	
	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest bartalis@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bartalis@tmit.bme.hu
Source	METANET4U	
Metadata language ID	en-us	
Metadata last date updated	2013-01-22	
Revision	2013-01-22	

Validation

Validated	True	
Type	Content	
Mode	Manual	
Details	Manually checked annotation and labeling of the sound and word boundaries in the corpora and visually checked the formant values in the vowels	
Validator	Tamás Gábor Csapó	
	Position	Ph.D. candidate
	Contact	Magyar tudósok körútja 2. H-1117 Budapest csapot@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics csapot@tmit.bme.hu

Usage

Foreseen use	Human use NLP applications
NLP-specific use	Knowledge discovery Linguistic research

	Speech analysis	
Actual uses	NLP applications	
	NLP-specific use	Speech analysis
	Derived resource	The first Hungarian interactive formant database
	Actual use details	http://beszedmuhely.tmit.bme.hu/formant

Resource creation

Resource creator	Csaba Zainkó	
	Position	Ph.D. lecturer
	Contact	Magyar tudósok körútja 2. H-1117 Budapest zainko@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics zainko@tmit.bme.hu
	Tamás Böhm	
	Position	Ph.D. lecturer
	Contact	Magyar tudósok körútja 2. H-1117 Budapest bohm@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bohm@tmit.bme.hu
Creation start date	2007-05-22	
Creation end date	2013-01-17	

Resource documentation

Reports	http://speechlab.tmit.bme.hu/CESAR/formant_hu_description_en.pdf http://speechlab.tmit.bme.hu/CESAR/formant_hu_description_hu.pdf
----------------	--

Texts

Media type	text
Linguality type	Monolingual
Languages	Hungarian

	Language ID	HU
	Size	3608 words
Size	3608 words	
Character encoding	ISO-8859-2	
Creation	Original source	research
	Creation mode	Manual

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	3318 seconds
Audio size	3318 seconds (3318 seconds of effective speech in 3318 seconds of audio content)	
Audio content	Speech items	Isolated words
	Noise level	Low
Setting	Naturality	Read speech
	Conversational type	Monologue
	Scenario type	Other
	Audience	No
	Interactivity	Non interactive
Audio formats	Wave/audio	
	Signal encoding	Linear PCM
	Sampling rate	22050
	Quantization	16
	Compression	False
	Number of tracks	1
	Recording quality	High
	Size	3318 seconds
Annotation	Segmentation	
	Annotated elements	Other
	Segmentation level	Phoneme

	Format	Praat TextGrid	
	Tagset	http://praat.org	
	Annotation mode	Mixed	
	Annotation mode details	TTS for the text to sound conversion, and manual alignment of sound boundaries	
	Annotation tool	Profivox development tool	
	Start date	2008-06-17	
	End date	2013-01-17	
	Size	32667 phonemes	
	Annotators	Klára Laczkó Position staff member Contact Magyar tudósok körútja 2. H-1117 Budapest klaklara@tmit.bme.hu http://www.tmit.bme.hu Organization Budapest University of Technology and Economics Department of Telecommunications and Media Informatics klaklara@tmit.bme.hu	
Recording	Device	Other	
	Device details	RME Fireface800	
	Platform software	soundforge_sony	
	Environment	Studio	
	Recorders	Mátyás Bartalis Position research fellow Contact Magyar tudósok körútja 2. H-1117 Budapest bartalis@tmit.bme.hu http://speechlab.tmit.bme.hu Organization Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bartalis@tmit.bme.hu	
Capture	Capturing device type	Studio equipment	
	Capturing device type details	AKG C-414 B-ULS	

	Person source set	Number of persons	2
		Age of persons	Adult
		Age range start	26
		Age range end	60
		Sex of persons	Mixed
		Origin of persons	Native
		Dialect accent of persons	no dialect
		Hearing impairment of persons	No
		Speaking impairment of persons	No
		Number of trained speakers	2
Creation	Original source	corpora	

2.19. Graphical query interface for Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection

General Information

Short name	F-view-hu
Description	This is a program for the visualization of the spectral shape and formant values (on 5 measured point in each vowel). This tool enables search in the phoneme sequences of the database by giving short phoneme sequences. The result gives those sentences where the short phoneme sequence occurs. The visualized formant values can be modified manually. The modification is stored automatically in the formant database. The tool is capable to study formant values and formant movements in different sound environment.
Identifier	219
Resource type	Tool/service
Tool/service type	Other
Version	v1.0
Last update	2013-01-22

Contacts

Géza Németh	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest nemeth@tmit.bme.hu http://speechlab.tmit.bme.hu
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Gábor Olaszy	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
Availability start date	2013-01-25	

Licences

CLARIN_RES		
Restrictions of use	Academic – non-commercial use	
Access medium	Hard disk	
Fee	5000 EURO	
Signatories	Henk Tamás	
	Position	Head of Department
	Contact	Magyar tudósok körútja 2. H-1117 Budapest henk@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics henk@tmit.bme.hu

Distribution rights holder	Gábor Olasz	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu

Metadata

Creation date	2013-01-22	
Metadata creators	Mátyás Bartalis	
	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest bartalis@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bartalis@tmit.bme.hu
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2013-01-22	

Validation

Validated	True	
Type	Content	
Mode	Manual	
Validator	Bálint Tóth	
	Position	Ph.D. candidate
	Contact	Magyar tudósok körútja 2. H-1117 Budapest toth.b@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics

	toth.b@tmit.bme.hu
--	--

Usage

Actual uses	Human use	
	Reports	Abari Kálmán, Olasz Gábor: Interaktív formánsérték-módosító fejlesztése, MSZNY 2011, pp. 309-315, http://www.inf.u-szeged.hu/mszny2011/images/stories/kepek/mszny2011_press_nc_b5.pdf

Resource creation

Resource creator	Gábor Olasz	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
	Kálmán Abari	
	Contact	abari.kalman@gmail.com
Creation start date	2008-01-01	
Creation end date	2013-01-20	

Resource documentation

Reports	http://speechlab.tmit.bme.hu/CESAR/tool_formant_hu_description_en.pdf http://speechlab.tmit.bme.hu/CESAR/tool_formant_hu_description_hu.pdf
----------------	--

Tool/service

Tool/service type	Other	
Tool/service subtype	search and visualization of formant data in the function of sound environment	
Language dependent	True	
Input	Media type	text
Output	Media type	image text
Operating system	Windows	
Required software	http://www.r-project.org/	
Required hardware	None	

Required LR s	none		
Tool/service evaluation	Evaluated	True	
	Level	Usage	
	Evaluators	Gábor Olaszy	
		Position	Professor
		Contact	olaszy@tmit.bme.hu

2.20. Hungarian Medical Speech Database

General Information

Short name	HuMedical
Description	Hungarian Medical Speech Database is a newly created and continuously enhanced speech database which contains pathological speech uttered by speakers suffering from various speech disorders. The development of the database was carried out within CESAR.
Identifier	220
Resource type	Corpus
URL	http://alpha.tmit.bme.hu/speech/
Version	1.0
Last update	2012-12-20

Contacts

Klára Vicsi	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest

		vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Availability start date	2013-01-31	

Licences

MS-NC-NoReD-FF		
Restrictions of use	Academic – non-commercial use	
Access medium	DVD-R	
Download location	http://alpha.tmit.bme.hu/speech/Medical.php	
Signatories	Henk Tamás	
	Position	Head of Department
	Contact	Magyar tudósok körútja 2. H-1117 Budapest henk@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics henk@tmit.bme.hu
Distribution rights holder	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu

Metadata

Creation date	2013-01-22	
Metadata creators	György Szaszák	
	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/speech/

	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2013-01-22	

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Other Speech analysis Speech understanding	
Actual uses	Human use	
	NLP-specific use	Speech analysis

Resource creation

Resource creator	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu
Creation start date	2011-01-01	
Creation end date	2012-12-20	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Size	50 files	

Character encoding	UTF-8	
Creation	Creation mode	Manual

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	2 hours
Audio size	2 hours	
Audio content	Speech items	Free speech Phonetically rich sentences
	Noise level	Low
Setting	Naturality	Assisted
	Conversational type	Dialogue
	Scenario type	Other
	Audience	Few
	Interactivity	Semi interactive
Audio formats	Wave/audio	
	Signal encoding	Linear PCM
	Sampling rate	16000
	Quantization	16
	Compression	False
	Number of tracks	1
	Recording quality	High
	Size	2 hours
Annotation	Speech annotation – orthographic transcription	
	Annotated elements	Background noise
	Segmentation level	Phrase
	Format	TextGrid (Praat)
	Annotation mode	Manual

	Annotation mode details	Segmentated into speaker turns, partial phoneme segmentation
	Annotation tool	Praat
	Start date	2011-05-01
	End date	2012-12-20

2.21. Automatic Prosodic Segmenter

General Information

Short name	ProSeg
Description	Automatic Prosodic Segmenter is a tool designed for prosodic segmentation of speech utterances down to the phonological phrase level if possible. The system is designed to be language independent, however, it highly relies on fixed stress, and hence, the range of other languages for which the system is adaptable can be restricted. Hungarian read speech phonological phrase models are provided, but the system is retrainable using the HTK toolkit for other languages provided that data is available.
Identifier	221
Resource type	Tool/service
Tool/service type	Tool
URL	http://alpha.tmit.bme.hu/speech/
Version	1.0
Last update	2012-12-20

Contacts

György Szaszák	
Position	researcher
Contact	Magyar tudósok körútja 2. H-1117 Budapest szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	György Szaszák	
	Position	researcher
	Contact	Magyar tudósok körútja 2.

		H-1117 Budapest szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu
Availability start date	2013-01-31	

Licences

GPL		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://alpha.tmit.bme.hu/speech/ProSeg.php	
Signatories	Henk Tamás	
	Position	Head of Department
	Contact	Magyar tudósok körútja 2. H-1117 Budapest henk@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics henk@tmit.bme.hu
Distribution rights holder	György Szaszák	
	Position	researcher
	Contact	Magyar tudósok körútja 2. H-1117 Budapest szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu

Metadata

Creation date	2013-01-22	
Metadata creators	György Szaszák	
	Position	researcher
	Contact	Magyar tudósok körútja 2. H-1117 Budapest

		szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2013-01-22	

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Information extraction Linguistic research Speech analysis	
Actual uses	Human use	
	NLP-specific use	Information extraction Natural language understanding Speech analysis

Resource creation

Resource creator	György Szaszák	
	Position	researcher
	Contact	Magyar tudósok körútja 2. H-1117 Budapest szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu
Creation start date	2006-01-01	
Creation end date	2009-06-30	

Tool/service

Tool/service type	Tool
Language dependent	True

2.22. Hungarian Phonetic Transcriber

General Information

Short name	PhonTrans
Description	Hungarian Phonetic Transcriber is a rule based application designed to perform automatic phonetic transcription. As written and spoken language are close in standard formal Hungarian, the tool is highly reliable, however, when intended to used for the transcription of proper nouns, a dictionary holding exceptions and special pronunciations is required. This is not included but can be compiled from other resources available in CESAR. The tool uses the SAMPA phonetic alphabet.
Identifier	222
Resource type	Tool/service
Tool/service type	Tool
URL	http://alpha.tmit.bme.hu/speech/
Version	1.0
Last update	2007-07-20

Contacts

Klára Vicsi	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu

Distribution

Availability	Available – restricted use	
IPR holder	Klára Vicsi	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics

	vicsi@tmit.bme.hu
Availability start date	2013-01-31

Licences

MS-NC-NoReD	
Restrictions of use	Academic – non-commercial use
Access medium	Downloadable
Download location	http://alpha.tmit.bme.hu/speech/PhonTrans.php
Signatories	Henk Tamás
	Position Head of Department
	Contact Magyar tudósok körútja 2. H-1117 Budapest henk@tmit.bme.hu http://www.tmit.bme.hu
	Organization Budapest University of Technology and Economics Department of Telecommunications and Media Informatics henk@tmit.bme.hu
Distribution rights holder	Klára Vicsi
	Position professor
	Contact Magyar tudósok körútja 2. H-1117 Budapest vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization Budapest University of Technology and Economics Department of Telecommunications and Media Informatics vicsi@tmit.bme.hu

Metadata

Creation date	2013-01-22
Metadata creators	György Szaszák
	Position research fellow
	Contact Magyar tudósok körútja 2. H-1117 Budapest szaszak@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization Budapest University of Technology and Economics Department of Telecommunications and Media Informatics szaszak@tmit.bme.hu

Source	CESAR
Metadata language ID	en-us
Metadata last date updated	2013-01-22

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Lexicon enhancement	
Actual uses	Human use	
	NLP-specific use	Lexicon enhancement

Resource creation

Resource creator	Szabolcs Velkei	
	Contact	velkei@synaptic.hu http://www.synaptic.hu
	Organization	Synaptic Ltd. velkei@synaptic.hu
Creation start date	2006-01-01	
Creation end date	2007-06-20	

Tool/service

Tool/service type	Tool
Language dependent	True

2.23. Hungarian MALACH Database

General Information

Description	The Hungarian MALACH Speech Database is created as a part of the MALACH (Multilingual Access to Large Spoken Archives) database. It contains 116,000 hours of digitized interviews in 32 languages from 52,000 survivors, liberators, rescuers and witnesses of the Nazi Holocaust. The presented hungarian database includes 24 whole interviews and parts of 104 interviews with Hungarian speakers. The recordings are segmented between speech pauses for speech recognition software development purposes. The interview parts should be used to train the recognizer, the complete interviews are chosen for testing. Manual transcriptions for all the audio files are also included in the database.
Identifier	223

Resource type	Corpus
Version	1.0

Contacts

Tibor Fegyó	
Contact	tfegy@aitia.ai http://www.tmit.bme.hu/fegy.tibor

Distribution

Availability	Available – restricted use	
IPR holder	USC Shoah Foundation Institute	
	Contact	yhi-web@usc.edu

Licences

Proprietary		
Access medium	DVD-R	
Attribution text	Mihajlik P, Tuske Z, Tarjan B, Nemeth B, Fegyó T Improved Recognition of Spontaneous Hungarian Speech.: Morphological and Acoustic Modeling Techniques for a Less Resourced Task. IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING 18:(6) pp. 1588-1600. Paper TASL.2009.2038807. (2010)	
Signatories	Fegyó Tibor	
	Position	Director of speech research
	Contact	tfegy@aitia.ai http://www.tmit.bme.hu/fegy.tibor
	Organization	AITIA International, Inc tfegy@aitia.ai

Metadata

Creation date	2013-01-23	
Metadata creators	Tibor Fegyó	
	Contact	tfegy@aitia.ai
Source	METANET4U	
Metadata language ID	en-us	
Metadata last date updated	2013-01-23	

Texts

--	--

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Size	1345000 phonemes	
Character encoding	windows-1250	
Creation	Creation mode	Manual
	Creation tools	Transcriber

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Audio size	19 gb (59 hours of audio content)	
Audio content	Speech items	Free speech
	Noise level	Low
Setting	Naturality	Spontaneous
	Conversational type	Dialogue
	Scenario type	Other
	Audience	No
Audio formats	Wave/audio	
	Signal encoding	Linear PCM
	Sampling rate	48000
	Quantization	16
	Compression	True
	Compression name	Mp3
	Compression loss	True
	Number of tracks	1
	Recording quality	High
Annotation	Speech annotation – orthographic transcription	
	Format	.mlf

	Annotation mode	Manual
	Annotation tool	Transcriber

2.24. Hungarian Broadcast Conversation Database from the Catholic Radio of Eger (EKR)

General Information

Description	The recordings are mostly complete radio programmes, and contain semi-spontaneous conversations. For every sound file there is a transcription text, which is created with the Transcriber tool manually.
Identifier	224
Resource type	Corpus
Version	1.0

Contacts

Tibor Fegyó	
Contact	tfegyo@aitia.ai http://www.tmit.bme.hu/fegyo.tibor

Distribution

Availability	Available – restricted use	
IPR holder	Szent István Rádió	
	Contact	info@szentistvanradio.hu

Licences

MS-C-NoReD-FF		
Access medium	DVD-R	
Fee	The audio corpus: 3000 Euro, the text corpus 15000 Euro	
Signatories	Fegyó Tibor	
	Position	Director of speech research
	Contact	tfegyo@aitia.ai http://www.tmit.bme.hu/fegyo.tibor
	Organization	AITIA International, Inc tfegyo@aitia.ai

Metadata

Creation date	2013-01-23
----------------------	------------

Metadata creators	Tibor Fegyó	
	Contact	tfegyó@aitia.ai
Source	METANET4U	
Metadata language ID	en-us	
Metadata last date updated	2013-01-23	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Size	3000000 phonemes	
Character encoding	windows-1250	
Creation	Creation mode	Manual
	Creation tools	Transcriber

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
Audio size	20 gb (66 hours of audio content)	
Audio content	Speech items	Free speech
	Noise level	Low
Setting	Naturality	Spontaneous
	Conversational type	Dialogue
	Scenario type	Other
	Audience	No
Audio formats	Wave/audio	
	Signal encoding	Linear PCM
	Sampling rate	44100
	Quantization	16
	Compression	True

	Compression name	Mp3
	Compression loss	True
	Number of tracks	1
	Recording quality	High
Annotation	Speech annotation – orthographic transcription	
	Format	.trs
	Annotation mode	Manual
	Annotation tool	Transcriber

2.25. Accent marker database for Hungarian written sentences

General Information

Short name	accent-hu
Description	This sentence corpus is supplied with yes/no accent markers on each word.
Identifier	225
Resource type	Corpus
Version	1.0
Revision	Texts with yes/no accent markers
Last update	2013-01-22

Contacts

Géza Németh	
Position	professor
Contact	Magyar tudósok körútja 2. H-1117 Budapest nemeth@tmit.bme.hu http://speechlab.tmit.bme.hu
Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics nemeth@tmit.bme.hu

Distribution

Availability	Available – restricted use

IPR holder	Gábor Olszy	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest olaszy@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics olaszy@tmit.bme.hu
Availability start date	2013-01-24	

Licences

CLARIN_RES		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://speechlab.tmit.bme.hu/CESAR/accent_hu.zip	
Fee	free of charge	
Signatories	Henk Tamás	
	Position	Head of Department
	Contact	Magyar tudósok körútja 2. H-1117 Budapest henk@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics henk@tmit.bme.hu
Distribution rights holder	Géza Németh	
	Position	professor
	Contact	Magyar tudósok körútja 2. H-1117 Budapest nemeth@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics nemeth@tmit.bme.hu

Metadata

Creation date	2013-01-22

Metadata creators	Mátyás Bartalis	
	Position	research fellow
	Contact	Magyar tudósok körútja 2. H-1117 Budapest bartalis@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bartalis@tmit.bme.hu
Source	METANET4U	
Metadata language ID	en-us	
Metadata last date updated	2013-01-22	
Revision	2013-01-22	

Validation

Validated	True	
Type	Content	
Mode	Manual	
Details	Manually checked accent markers on words	
Validator	Tamás Gábor Csapó	
	Position	Ph.D. candidate
	Contact	Magyar tudósok körútja 2. H-1117 Budapest csapot@tmit.bme.hu http://www.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics csapot@tmit.bme.hu

Usage

Foreseen use	Human use NLP applications
NLP-specific use	Knowledge discovery Linguistic research Tex to speech synthesis

Resource creation

--	--

Resource creator	Csaba Zainkó	
	Position	Ph.D. lecturer
	Contact	Magyar tudósok körútja 2. H-1117 Budapest zainko@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics zainko@tmit.bme.hu
	Tamás Bőhm	
	Position	Ph.D. lecturer
	Contact	Magyar tudósok körútja 2. H-1117 Budapest bohm@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics bohm@tmit.bme.hu
Creation start date	2012-06-11	
Creation end date	2013-01-17	

Resource documentation

Reports	http://speechlab.tmit.bme.hu/CESAR/accent_hu_description_en.pdf http://speechlab.tmit.bme.hu/CESAR/accent_hu_description_hu.pdf
----------------	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	1950 sentences
Size	1950 sentences	
Character encoding	ISO-8859-2	
Creation	Original source	research
	Creation mode	Manual

3. FFZG resources

3.1. Croatian National Corpus v2.5

General Information

Short name	HNK
Description	The Croatian National Corpus (HNK) is a representative corpus of contemporary Croatian standard language written texts published since 1990. The corpus is automatically lemmatised and MSD tagged. The documents are annotated with their genre, type and other information. The whole corpus is composed of fiction, non-fiction and mixed texts. This is a pseudocorpus, only the query interface using Bonito client is available, while the original texts cannot be distributed for copyright reasons. Bonito client gives opportunities for issue complex queries due to elaborated query language resulting not only in concordances, but also in word-lists, collocations and other types of distributional data etc. of tokens, lemmas and/or MSDs.
Identifier	301
Resource type	Corpus
URL	http://hmk.ffzg.hr/ http://filip.ffzg.hr/bonito2/run.cgi/first_form?corpname=HNK_v25
Version	2.5.1
Last update	2011-09-01

Contacts

Marko Tadić	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics

Distribution

Availability	Available – restricted use	
IPR holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Licences

Proprietary		
Restrictions of use	Academic – non-commercial use	
Access medium	Accessible through interface	
Execution location	http://hmk2.ffzg.hr	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Metadata

Creation date	2011-11-26	
Metadata creators	Marko Tadić	
	Position	Head of the Chair for Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
Metadata language ID	en	
Metadata last date updated	2013-02-06	

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net

	Funding type	EU funds National funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	1998-12-01	

Resource documentation

Reports	<p>Marko Tadić. Building the Croatian National Corpus. LREC2002 Proceedings, Las Palmas-Pariz, 2002, Vol. II, 2002, pp 441-446</p> <p>Marko Tadić. Developing the Croatian National Corpus and Beyond. Grzybek, Peter (ed.) Contributions to the Science of Text and Language. Word Length Studies and Related Issues, Kluwer, Dordrecht 2006, pp 295-300</p> <p>Marko Tadić. New version of the Croatian National Corpus. Hlaváčková, Dana; Horák, Aleš; Osolsobě, Klara; Rychlý, Pavel (eds.) After Half a Century of Slavonic Natural Language Processing, Masaryk University, Brno, 2009, pp 199-205</p>
----------------	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	Latn
Size	101000000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word

Morphosyntactic annotation – b pos tagging	
Segmentation level	Word

3.2. Croatian Morphological Lexicon v4.6

General Information

Short name	HML
Description	The Croatian Morphological Lexicon is an inflectional lexicon generated automatically by Croatian Inflectional Generator from ca 110,000 lemmas yielding over 4,000,000 word forms. It has been a result of the group lead by Prof. Marko Tadić on the basis of theoretical background published in 1992 (see Tadić 1994 above). The initial set of lemmas was collected from several existing Croatian mono- and bi-lingual dictionaries, while additional entries were collected via corpus or by means of automatic enlargement of the initial list of lemmas (see Bekavac, Šojat 2005, and Oliver, Tadić 2004 above). The automatically generated output was corrected for known systemic errors, encoded in utf-8 and stored in MulTextEast Lexica format: lemma[TAB]word-form[TAB]MSD. The MSD-tagset is conformant with the MulTextEast v4.0 recommendations for Croatian language. However, some additions exist: in surnames gender is left unspecified (-), additional subclassification of adverbials has been introduced etc. At the moment the Croatian Morphological Lexicon is a pseudolexicon, accessible only through the Croatian Lemmatisation Server web query interface or php script call.
Identifier	302
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
Version	4.6

Contacts

Marko Tadić	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics

Distribution

Availability	Available – restricted use	
IPR holder	Marko Tadić	
	Contact	Ivana Lučića 3 10000 Zagreb

		marko.tadic@ffzg.hr http://hmk.ffzg.hr
	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Licences

Proprietary		
Restrictions of use	Academic – non-commercial use	
Access medium	Accessible through interface	
Execution location	http://hmk.ffzg.hr	
Fee	negotiable	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Metadata

Creation date	2011-11-26	
Metadata creators	Marko Tadić	
	Position	Head of the Chair of Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
Metadata language ID	en	
Metadata last date updated	2013-02-01	

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzk@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2003-04-01	

Resource documentation

Reports	<p>Marko Tadić. Računalna obradba morfologije hrvatskoga književnoga jezika. PhD Thesis, University of Zagreb, Faculty of Humanities and Social Sciences, 1994.</p> <p>Marko Tadić, Sanja Fulgosi. Building the Croatian Morphological Lexicon. Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages, ACL, Budapest, 2003, pp 41-46</p> <p>Antoni Oliver, Marko Tadić. Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora. LREC2004 Proceedings, Lisbon-Pariz, 2004, Vol. IV, pp 1259-1262</p> <p>Bekavac, Božo; Šojat, Krešimir. Lexical acquisition through particular adjectival endings for Croatian. Workshop on Computational Modeling of Lexical Acquisition, Split, 2005.</p>
----------------	---

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology
	Linguistic information	Lemma Lemma – abbreviations Auxiliary

		Case Degree Gender Inflection Mood Number Person Tense Other Part of speech
Creation	Original source	Headword lists from different Croatian monolingual dictionaries, Croatian National Corpus, Croatian Web-Corpus
	Creation mode	Mixed
	Creation tools	Croatian Word-Forms Generator

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	Latn
Size	4000000 entries	
Character encoding	UTF-8	

3.3. Croatian-English Parallel Corpus

General Information

Short name	Hr-En p-corp
Description	The Croatian-English Parallel Corpus (Hr-En p-corp) is a parallel unidirectional (hr to en) corpus of contemporary Croatian standard language collected from articles appearing in Croatia Weekly newspapers, published from 1998 to 2000. The corpus samples were obtained in digital form entirely, converted to XML, aligned using Vanilla Aligner, manually checked and stored in TMX format.
Identifier	303
Resource type	Corpus
URL	http://hmk.ffzg.hr/hr-en_p-corp
Version	2
Last update	2011-09-01

Contacts

Marko Tadić	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics

Distribution

Availability	Available – unrestricted use	
IPR holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Licences

CC-BY-NC-SA		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://hmk.ffzg.hr/hr-en_p-corp/download/CW_v02_tmx.zip	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Metadata

Creation date	2011-11-26	
Metadata creators	Marko Tadić	

	Position	Head of the Chair for Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
Metadata language ID	en	
Metadata last date updated	2011-11-27	

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2000-03-01	

Resource documentation

Reports	Marko Tadić. Building the Croatian-English Parallel Corpus. LREC2000 Proceedings, Athens-Pariz, 2000, Vol. I, pp 523-530 Marko Tadić. Procedures in Building the Croatian-English Parallel Corpus. International Journal of Corpus Linguistics, special issue, (2001), pp 107-123
----------------	--

Texts

Media type	text

Linguality type	Bilingual	
Languages	Croatian	
	Language ID	hr
	Language script	Latn
	English	
	Language ID	en
	Language script	Latn
Size	62500 units	
Character encoding	UTF-8	
Annotation	Alignment	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Sentence

3.4. Croatian Lemmatisation Server

General Information

Short name	CLS
Description	The Croatian Lemmatisation Server (CLS) is a web-based service for lemmatisation, POS- and MSD-tagging of Croatian texts. It accepts input in two modes. Through web form mode it accepts direct query allowing lemmas or word-forms as input, giving all word-forms of lemma or all lemmas that a word-form could belong to, respectively. In both cases, the results are accompanied by MSD-tags as well. In the upload mode the CLS expects a verticalised, utf-8 encoded text in contemporary standard Croatian language and returns a zip file with results of processing the uploaded file. At the moment the limitation of file size is 50,000 tokens. The processing gives all analysis for each token, i.e. line in verticalised corpus, regarding the lemma, POS and MSD. The web interface allows user to select the level of processing needed: just lemmatisation, lemmatisation with POS-tagging or lemmatisation with MSD-tagging. POS and MSD tags follow the MulTextEast v4.0 specifications for Croatian. Upon registration either as academic or commercial user, a php script call tailored according to user's requests can be provided. Also, the existing Croatian Lemmatisation Server will be turned into a web service that will feature lemmatisation and MSD-tagging of verticalised utf-8 encoded Croatian texts including disambiguation.
Identifier	304
Resource type	Tool/service
Tool/service type	Service
URL	http://hml.ffzg.hr/
Version	2.0

Last update	2011-11-20
--------------------	------------

Contacts

Marko Tadić	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics

Distribution

Availability	Available – restricted use	
IPR holder	Marko Tadić	
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr
	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Licences

Proprietary		
Restrictions of use	Commercial use	
Access medium	Accessible through interface	
Execution location	http://hmk.ffzg.hr	
Fee	negotiable	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3

		10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr
--	--	---

Metadata

Creation date	2011-11-20	
Metadata creators	Marko Tadić	
	Position	Head of the Chair of Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2005-04-01	

Resource documentation

Reports	Marko Tadić. The Croatian Lemmatization Server. Southern Journal of Linguistics 29 (2005), 1/2, pp 206-217 Marko Tadić. Croatian Lemmatization Server. Mila Dimitrova Vulchanova, Svetla Koeva, Iliyana Krapova, Valentin Vulchanov (eds.). Formal Approaches to south Slavic and Balkan Languages, Bulgarian Academy of Sciences, Sofia, 2006, pp 140-146
----------------	---

Tool/service

Tool/service type	Service	
Language dependent	True	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	Linux	
Required software	php	
Tool/service evaluation	Evaluated	True
	Level	Diagnostic
	Type	Black box
	Criteria	Intrinsic
	Measure	Human
	Evaluators	Marko Tadić
		Contact Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hnk.ffzg.hr
		Organization University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences
Tool/service creation	Implementation language	php

3.5. Croatian Valency Lexicon

General Information

Short name	CROVALLEX
Description	The Croatian Valency Lexicon of Verbs, Version 2.0008 (CROVALLEX 2.0008) is an attempt of formal description of valency frames of Croatian verbs. CROVALLEX 2.0008 was developed as the part of the PhD thesis titled Approaches to the Development of the

	Machine Lexicon for Croatian Language written by Nives Mikelić Preradović and supervised by prof.dr.sc. Damir Boras at the Department of Information Sciences, Faculty of Humanities and Social Sciences, Zagreb University. The Functional Generative Description (FGD), being developed by Czech linguists Petr Sgall and his collaborators since the 1960s, is used as the background theory in CROVALLEX 2.0008. for the description of valency frames of selected verbs. CROVALLEX 2.0008 contains roughly 1740 verbs. They were selected from the Croatian frequency dictionary, according to their number of occurrences. The preparation of this version of CROVALLEX has taken around three years
Identifier	305
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://theta.ffzg.hr/crovallex/
Version	2.0008

Contacts

Nives Mikelić Preradović	
Position	Professor Assistant
Contact	Ivana Lučića 3 10000 Zagreb nmikelic@ffzg.hr http://www.ffzg.unizg.hr/infoz/hr/index.php/lanovi-odsjeaka/190-nives-mikeli-preradovi-
Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences

Distribution

Availability	Available – restricted use	
IPR holder	Nives Mikelić Preradović	
	Contact	Ivana Lučića 3 10000 Zagreb nmikelic@ffzg.hr http://www.ffzg.unizg.hr/infoz/hr/index.php/lanovi-odsjeaka/190-nives-mikeli-preradovi-
	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences	
	Contact	Ivana Lučića 3 10000 Zagreb npetak@ffzg.hr http://www.ffzg.unizg.hr/infoz/hr/index.php/odsjek

Licences

CC-BY-NC-SA

Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://theta.ffzg.hr/crovallex/	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences	
	Contact	Ivana Lučića 3 10000 Zagreb npetak@ffzg.hr http://www.ffzg.unizg.hr/infoz/hr/index.php/odsjek

Metadata

Creation date	2011-11-26	
Metadata creators	Marko Tadić	
	Position	Head of the Chair for Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department/Institute of Linguistics
Metadata language ID	en	
Metadata last date updated	2013-02-04	

Resource creation

Resource creator	Nives Mikelić Preradović	
	Contact	Ivana Lučića 3 10000 Zagreb nmikelic@ffzg.hr http://www.ffzg.unizg.hr/infoz/hr/index.php/lanovi-odsjeka/190-nives-mikeli-preradovi-
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)

	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2008-06-01	

Resource documentation

Reports	<p>Mikelić Preradović, Nives. Approaches to the Development of the Machine Lexicon for Croatian Language. PhD Thesis, University of Zagreb, Faculty of Humanities and Social Sciences, 2008.</p> <p>Mikelic Preradovic, Nives; Boras, Damir; Kišiček, Sanja. CROVALLEX: Croatian Verb Valence Lexicon. In: Lužar-Stiffler, Vesna ; Jarec, Iva ; Bekić, Zoran (eds.) Proceedings of the 31st International Conference on Information Technology Interfaces (ITI 2009), Zagreb : SRCE, 2009. pp. 533-538</p> <p>Mikelić Preradović, Nives. Semantic classification of verbs in CROVALLEX. In: Lagakos, Stephen ; Perlovsky, Leonid ; Jha, Manoj ; Covaci, Brindusa ; Zaharin, Azama ; Mastorakis, Nikos (eds.) Recent Advances in Computer Engineering and Applications. Proceedings of the 4th WSEAS International Conference on Computer Engineering and Applications (CEA '10). Harvard University, Cambridge, USA : WSEAS Press, 2010. pp. 53-59</p>
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Semantics Syntax
	Linguistic information	Lemma Other Part of speech Semantics – semantic roles
Creation	Original source	Headword list selected from Moguš, Milan ; Bratanić, Maja ; Tadić, Marko. Hrvatski čestotni rječnik (Croatian Frequency Dictionary), Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu - Školska knjiga, Zagreb, 1999.
	Creation mode	Mixed

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	Latn
Size	1740 entries	

Character encoding	UTF-8
---------------------------	-------

3.6. Croatian Web Corpus

General Information

Short name	hrWaC
Description	Croatian Web Corpus (hrWaC) is the largest collected corpus for Croatian so far. It was collected in 2011-06 by crawling the whole .hr internet domain yielding ca 1.2 billion tokens. The corpus has been cleaned of HTML code, lemmatised and MSD-tagged automatically using CroTag system (Agić et al., 2008). The compilation of the corpus is described in the TSD2011 paper Ljubešić, N., Erjavec, T. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. The morphosyntactically annotated and lemmatized corpus is distributed under the CC-BY-SA licence. It has been installed also in NoSketchEngine for free on-line querying: http://faust.ffzg.hr/bonito2/run.cgi/first_form?corpname=hrwac .
Identifier	306
Resource type	Corpus
URL	http://www.nljubescic.net/resources/corpora/hrwac/
Version	1.0
Last update	2012-07-30

Contacts

Nikola Ljubešić	
Position	Assistant Professor
Contact	Ivana Lučića 3 10000 Zagreb nljubesi@ffzg.hr http://www.nljubescic.net/
Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences nljubesi@ffzg.hr

Distribution

Availability	Available – restricted use	
IPR holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics, Department of Information Sciences
	Contact	Ivana Lučića 3 10000 Zagreb

		zzl@ffzg.hr http://hmk.ffzg.hr/
--	--	---

Licences

CC-BY-SA		
Restrictions of use	Attribution Share alike	
Access medium	Downloadable	
Execution location	http://www.nljubecic.net/resources/corpora/hrwac/	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Metadata

Creation date	2012-07-30	
Metadata creators	Marko Tadić	
	Position	Head of the Chair for Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics zzl@ffzg.hr
Metadata language ID	en	
Metadata last date updated	2013-02-04	

Resource creation

Resource creator	Univ. of Zagreb, Faculty of Humanities and Social Sciences, Depts. of Linguistics & Information Sci.	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short	CESAR

	name	
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2011-06-01	

Resource documentation

Reports	Nikola Ljubešić and Tomaž Erjavec. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. Text, Speech and Dialogue 2011. Lecture Notes in Computer Science, Springer.
----------------	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	Latn
Size	1 200 000 000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

3.7. Slovene Web Corpus

General Information

Short name	slWaC
Description	Slovene Web Corpus (slWaC) is the the first version of the Slovene web corpus. It was collected by crawling the whole .si internet domain in 2011-06 yielding ca 380 million tokens. The corpus has been lemmatised and MSD-tagged automatically using ToTaLe system (Erjavec et al. 2005). The compilation of the corpus is described in the TSD2011 paper Ljubešić, N., Erjavec, T. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. The morphosyntactically annotated and lemmatized corpus is distributed under the CC-BY-SA licence. The first version is freely accessible for querying at http://faust.ffzg.hr/bonito2/run.cgi/first_form?corpname=slwac . A new crawl with an updated crawler is scheduled for 2012-09. The target size of the second version of slWaC is 1 billion words.
Identifier	307
Resource type	Corpus
URL	http://www.nljubescic.net/resources/corpora/slwaC/
Version	1.0
Last update	2012-07-30

Contacts

Nikola Ljubešić	
Position	Assistant Professor
Contact	Ivana Lučića 3 10000 Zagreb nljubesci@ffzg.hr http://www.nljubescic.net/
Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences nljubesci@ffzg.hr

Distribution

Availability	Available – restricted use	
IPR holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics, Department of Information Sciences
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Licences

CC-BY-SA		
Restrictions of use	Attribution Share alike	
Access medium	Downloadable	
Execution location	http://www.nljubasic.net/resources/corpora/slwac/	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Metadata

Creation date	2012-07-30	
Metadata creators	Marko Tadić	
	Position	Head of the Chair for Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics zzl@ffzg.hr
Metadata language ID	en	
Metadata last date updated	2013-02-04	

Resource creation

Resource creator	Univ. of Zagreb, Faculty of Humanities and Social Sciences, Depts. of Linguistics & Information Sci.	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net

	Funding type	EU funds National funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2011-06-01	

Resource documentation

Reports	Nikola Ljubešić and Tomaž Erjavec. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. Text, Speech and Dialogue 2011. Lecture Notes in Computer Science, Springer.
----------------	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovenian	
	Language ID	sl
	Language script	Latn
Size	380 000 000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

3.8. Croatian-English Parallel Web Corpus

General Information

--	--

Short name	hrenWaC
Description	Croatian-English Parallel Web Corpus is a collection of paraellel Croatian-English texts crawled from .hr domain. This corpus was automatically collected by finding on-line documents in English that parallel to the documents already crawled in hrWaC. The parallelity of texts was calculated and selection treshold empirically set to 0.52 on a scale between 0 and 1. After that, the collection of parallel-text candidates has been manually inspected for real parallel texts. The initial crawled corpus had ca 253,000 sentence/translation units pairs (ca 8 Mw per language), while the manual checking resulted in 99,001 sentence/translation units pairs. The corpus is distributed under the CC-BY-SA licence.
Identifier	308
Resource type	Corpus
URL	http://www.nljubescic.net/resources/corpora/hrenwac/
Version	1.0
Last update	2012-07-30

Contacts

Nikola Ljubešić	
Position	Assistant Professor
Contact	Ivana Lučića 3 10000 Zagreb nljubesci@ffzg.hr http://www.nljubescic.net/
Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences nljubesci@ffzg.hr

Distribution

Availability	Available – restricted use	
IPR holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics, Department of Information Sciences
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Licences

CC-BY-SA

Restrictions of use	Attribution Share alike	
Access medium	Downloadable	
Execution location	http://www.nljubescic.net/resources/corpora/hrenwac/	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Metadata

Creation date	2012-07-30	
Metadata creators	Marko Tadić	
	Position	Head of the Chair for Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics zzl@ffzg.hr
Metadata language ID	en	
Metadata last date updated	2013-02-04	

Resource creation

Resource creator	Univ. of Zagreb, Faculty of Humanities and Social Sciences, Depts. of Linguistics & Information Sci.	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds
	Funder	European Commission (50%)

		University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-04-01	

Texts

Media type	text	
Linguality type	Bilingual	
Languages	Croatian	
	Language ID	hr
	Language script	Latn
	English	
	Language ID	en
	Language script	Latn
Size	99 001 units	
Character encoding	UTF-8	
Annotation	Alignment	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Sentence

3.9. South-East European Parallel Corpus

General Information

Short name	SETimes Corpus
Description	SouthEast European Parallel Corpus (SETimes Corpus) is based on the content published on the SETimes.com news portal. The news portal publishes “news and views from Southeast Europe” in ten languages: Albanian, Bosnian, Bulgarian, Croatian, English, Greek, Macedonian, Romanian, Serbian and Turkish. This version of the corpus tries to solve the issues present in an older version of the corpus (published inside OPUS, described in the LREC 2010 paper by Francis M. Tyers and Murat Serdar Alperen). The sentence-aligned language combinations are freely downloadable in TMX or TXT/Moses format. The corpus is published under the CC-BY-SA license.
Identifier	309
Resource type	Corpus

URL	http://www.nljubesic.net/resources/corpora/setimes/
Version	1.0
Last update	2012-07-30

Contacts

Nikola Ljubešić	
Position	Assistant Professor
Contact	Ivana Lučića 3 10000 Zagreb nljubesi@ffzg.hr http://www.nljubesic.net/
Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences nljubesi@ffzg.hr

Distribution

Availability	Available – restricted use	
IPR holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics, Department of Information Sciences
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Licences

CC-BY-SA		
Restrictions of use	Attribution Share alike	
Access medium	Downloadable	
Execution location	http://www.nljubesic.net/resources/corpora/setimes/	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Metadata

Creation date	2012-07-30	
Metadata creators	Marko Tadić	
	Position	Head of the Chair for Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics zzl@ffzg.hr
Metadata language ID	en	
Metadata last date updated	2013-02-04	

Resource creation

Resource creator	Univ. of Zagreb, Faculty of Humanities and Social Sciences, Depts. of Linguistics & Information Sci.	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-04-01	

Texts

Media type	text
Linguality type	Multilingual
Languages	Albanian

Language ID	sq
Language script	Latn
Bosnian	
Language ID	bs
Language script	Latn
Bulgarian	
Language ID	bg
Language script	Cyrl
Croatian	
Language ID	hr
Language script	Latn
English	
Language ID	en
Language script	Latn
Greek	
Language ID	el
Language script	GreK
Macedonian	
Language ID	mk
Language script	Cyrl
Romanian	
Language ID	ro
Language script	Latn
Serbian	
Language ID	sr
Language script	Cyrl
Turkish	
Language ID	tr
Language script	Latn

Size	43 142 458 tokens	
Character encoding	UTF-8	
Annotation	Alignment	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Sentence

3.10. Croatian Dependency Treebank

General Information

Short name	HOBS
Description	Croatian Dependency Treebank is a part of the Croatian National Corpus (i.e. Croatian part of the Croatian-English Parallel Corpus, CW2000) where 4,626 sentences (118,529 tokens) are planned to be manually annotated at the analytical layer following the Prague Dependency Treebank formalism adapted to Croatian. The corpus size is currently 3,465 sentences (88,045 tokens). It is published under CC-BY-NC-SA license.
Identifier	310
Resource type	Corpus
URL	http://hobs.ffzg.hr/
Version	1.0
Last update	2012-07-30

Contacts

Marko Tadić	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics

Distribution

Availability	Available – restricted use	
IPR holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department	Department/Institute of Linguistics

	name	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Licences

CC-BY-NC-SA		
Restrictions of use	Academic – non-commercial use Attribution Share alike	
Access medium	Downloadable	
Execution location	http://hobs.ffzg.hr	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Metadata

Creation date	2012-07-30	
Metadata creators	Marko Tadić	
	Position	Head of the Chair for Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
Metadata language ID	en	
Metadata last date updated	2013-02-04	

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department	Department/Institute of Linguistics

	name	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2007-06-01	

Resource documentation

Reports	Tadić, Marko. Building the Croatian Dependency Treebank: the initial stages. // <i>Suvremena lingvistika</i> . 33 (2007), 63; 85-92. Agić, Željko. Pristupi ovisnosnom parsanju hrvatskih tekstova / PhD thesis. Zagreb : University of Zagreb, Faculty of Humanities and Social Sciences, 2012-07-09, 216 p.
----------------	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	Latn
Size	88 045 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	

	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation – b pos tagging	
	Segmentation level	Word
	Syntactic annotation – treebanks	
	Segmentation level	Word

3.11. Web Content Extractor

General Information

Short name	WebContentExtractor
Description	Web Content Extractor is a tool for content extraction from web pages for building web corpora. The content extraction algorithm developed for building hrWaC and slWaC is described in TSD2011 paper Ljubešić, N., Erjavec, T. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. An implementation (a java file) is published under the Apache 2.0 licence. A Croatian evaluation sample used in the paper can also be downloaded and it is distributed under the CC-BY-SA license.
Identifier	311
Resource type	Tool/service
Tool/service type	Tool
URL	http://www.nljubestic.net/resources/tools/webcontentextractor/
Version	1.0
Last update	2012-07-30

Contacts

Nikola Ljubešić	
Position	Assistant Professor
Contact	Ivana Lučića 3 10000 Zagreb nljubesi@ffzg.hr http://www.nljubestic.net/
Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences nljubesi@ffzg.hr

Distribution

Availability	Available – unrestricted use	
IPR holder	Marko Tadić	
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr
	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics, Department of Information Sciences
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Licences

ApacheLicence_2.0		
Restrictions of use	Inform licensor	
Access medium	Downloadable	
Execution location	http://www.nljubescic.net/resources/tools/webcontentextractor/	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Metadata

Creation date	2012-07-30	
Metadata creators	Marko Tadić	
	Position	Head of the Chair for Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics zzl@ffzg.hr

Metadata language ID	en
Metadata last date updated	2013-02-04

Resource creation

Resource creator	Univ. of Zagreb, Faculty of Humanities and Social Sciences, Depts. of Linguistics & Information Sci.	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hnk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2011-04-01	

Tool/service

Tool/service type	Tool	
Language dependent	False	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	Linux	
Required software	Python (version 2.6 or higher)	
Tool/service evaluation	Evaluated	True
	Level	Diagnostic
	Type	Black box
	Criteria	Intrinsic

	Measure	Human	
	Evaluators	Nikola Ljubešić	
		Contact	Ivana Lučića 3 10000 Zagreb nljubesi@ffzg.hr http://www.nljubestic.net/
		Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences nljubesi@ffzg.hr
Tool/service creation	Implementation language	Python	

3.12. Collocation and Term Extractor

General Information

Short name	CollTerm
Description	CollTerm is a language independent tool for collocation and term extraction. It is an application that collects collocation and term candidates based on five different co occurrence measures for multiword units (i.e. collocations) or distributional differences from large representative corpus by application of the TF-IDF measurement on singleword units. The language dependent part consists of stop-word list and list of MWU MSD-patterns that can be coded with regular expressions as well. The application is describe in the paper presented at TKE2012 by Pinnis, M., Ljubešić, N., Ștefănescu, D., Skadiņa, I, Tadić, Gornostay, T. Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. The first version of this application is available as an integral part of ACCURAT Toolkit that is available under Apache 2.0 license (http://www accurat-project.eu/index.php?p=accurat-toolkit). In this version of the tool a calibration of MWU MSD-patterns has been provided for Croatian thus enhancing the usability of the tool. The plan is to provide calibration for other CESAR languages as well.
Identifier	312
Resource type	Tool/service
Tool/service type	Tool
URL	http://www.nljubestic.net/resources/tools/collterm/
Version	1.0
Last update	2012-07-30

Contacts

Nikola Ljubešić	
Position	Assistant Professor
Contact	Ivana Lučića 3 10000 Zagreb

	nljubesi@ffzg.hr http://www.nljubestic.net/
Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences nljubesi@ffzg.hr

Distribution

Availability	Available – unrestricted use	
IPR holder	Marko Tadić	
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr
	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics, Department of Information Sciences
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Licences

ApacheLicence_2.0		
Restrictions of use	Inform licensor	
Access medium	Downloadable	
Execution location	http://www.nljubestic.net/resources/tools/colterm/	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Metadata

Creation date	2012-07-30	
Metadata creators	Marko Tadić	
	Position	Head of the Chair for Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3

		10000 Zagreb marko.tadic@ffzg.hr
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics zzl@ffzg.hr
Metadata language ID	en	
Metadata last date updated	2013-02-04	

Resource creation

Resource creator	Univ. of Zagreb, Faculty of Humanities and Social Sciences, Depts. of Linguistics & Information Sci.	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
	Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation	
	Project short name	ACCURAT
	URL	http://www accurat-project.eu
	Funding type	EU funds National funds
	Funder	European Commission (75%) University of Zagreb, Faculty of Humanities and Social Sciences (25%)
	Start date	2010-01-01
	End date	2012-06-30
Creation start date	2011-04-01	

Tool/service

Tool/service type	Tool	
Language dependent	False	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	Linux	
Required software	Python (version 2.6 or higher)	
Tool/service evaluation	Evaluated	True
	Level	Diagnostic
	Type	Black box
	Criteria	Intrinsic
	Measure	Human
	Evaluators	Nikola Ljubešić
		Contact Ivana Lučića 3 10000 Zagreb nljubesi@ffzg.hr http://www.nljubesic.net/
		Organization University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences nljubesi@ffzg.hr
Tool/service creation	Implementation language	Python

3.13. Croatian Language Web Services

General Information

Short name	hrWS
Description	The Croatian Language Web Services (hrWS) is a set of language processing web services oriented towards the processing of Croatian language by evoking different modules. The modules supported with this version of hrWS are: sentence splitting, tokenization, PoS/MSD-tagging, NERC, dependency parsing. The hrWS use standard REST protocol for input and output of data to process.
Identifier	313
Resource type	Tool/service
Tool/service type	Tool

URL	http://lt.ffzg.hr/
Version	3.0
Last update	2013-01-30

Contacts

Željko Agić	
Position	senior research assistant
Contact	Ivana Lučića 3 10000 Zagreb zagic@ffzg.hr http://hmk.ffzg.hr/mt
Organization	Univ. of Zagreb, Faculty of Humanities and Social Sciences, Dept. of Information Sciences

Distribution

Availability	Available – restricted use	
IPR holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Licences

Restrictions of use	Commercial use	
Access medium	Accessible through interface	
Execution location	http://lt.ffzg.hr	
Fee	negotiable	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Metadata

--	--

Creation date	2013-01-30
Metadata creators	Marko Tadić
	Position Head of the Chair of Algebraic and Computational Linguistics
	Contact Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hnk.ffzg.hr/mt
	Organization University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
Metadata language ID	en
Metadata last date updated	2013-02-04

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences
	Short name FFZG
	Department name Department/Institute of Linguistics
	Contact Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hnk.ffzg.hr
Funding projects	Central and South-East European Resources
	Project short name CESAR
	URL http://www.cesar-project.net
	Funding type EU funds National funds Own funds
	Funder European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date 2011-02-01
	End date 2013-01-31
Creation start date	2011-06-01

Resource documentation

Reports	Marko Tadić. Hrvatski jezičnotehnoški web-servisi, 2013, DJT2013, Zagreb, 2012-11-30; http://www.cesar-project.net/events/croatia_roadshow/tadic.pptx
----------------	---

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	OS-independent	
Tool/service creation	Implementation language	C++ Java Python Perl

3.14. Croatian Translations of Acquis

General Information

Short name	hrAcquis
Description	The Croatian Translations of Acquis Communautaire is a bilingual English-Croatian parallel corpus that is following the specification set out by the JRC-Acquis Parallel Corpus. This corpus composed of 16638 documents translated within the Translation Service of the Ministry of Foreign Affairs and European Integrations of the Republic of Croatia until 2012-08-28. The conversion from .doc to XML was following the JRC-Acquis DTD. The English part of the Corpus was extracted from the original JRC-Acquis Corpus using CELEX codes to filter out the needed documents. The English-Croatian sentence alignment was processed with HunAlign alignment tool. The corpus is distributed under the CC-BY-SA licence.
Identifier	314
Resource type	Corpus
URL	http://meta-share.ffzg.hr/repository/hrAcquis
Version	1.0
Last update	2013-01-30

Contacts

Marko Tadić	
Position	Head of the Chair for Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt

Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Institute/Department of Linguistics marko.tadic@ffzg.hr
---------------------	--

Distribution

Availability	Available – restricted use	
IPR holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Licences

CC-BY-SA		
Restrictions of use	Attribution	
	Share alike	
Access medium	Downloadable	
Execution location	http://meta-share.ffzg.hr/repository/hrAcquis/	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Metadata

Creation date	2013-01-30	
Metadata creators	Marko Tadić	
	Position	Head of the Chair for Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics zzl@ffzg.hr
Metadata language ID	en	

Metadata last date updated	2013-01-30
-----------------------------------	------------

Resource creation

Resource creator	Univ. of Zagreb, Faculty of Humanities and Social Sciences, Institute/Department of Linguistics	
	Contact	Ivana Lučića 3 10000 Zagreb zzk@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-06-01	

Texts

Media type	text	
Linguality type	Bilingual	
Languages	Croatian	
	Language ID	hr
	Language script	latin
	English	
	Language ID	en
	Language script	latin
Size	30486240 tokens, 684238 units	
Character encoding	UTF-8	
Annotation	Alignment	
	Segmentation level	Sentence

	Segmentation	
	Segmentation level	Sentence

3.15. Croatian National Corpus v3.0

General Information

Short name	HNK v3.0
Description	The Croatian National Corpus (HNK) is a representative corpus of contemporary Croatian standard language written texts published since 1990. The corpus is automatically lemmatised and MSD tagged. The documents are annotated with their genre, type and other information. The whole corpus is composed of fiction, fiction and mixed texts. This is a pseudocorpus, only the query interface using Bonito2 web interface is available, while the original texts cannot be distributed for copyright reasons. Bonito2 web interface gives opportunities to issue complex queries due to elaborated query language resulting not only in concordances, but also in word-lists, collocations and other types of distributional data etc. of tokens, lemmas and/or MSDs. This version of HNK features Bonito2 web interface and additional texts
Identifier	315
Resource type	Corpus
URL	http://hnk.ffzg.hr/ http://filip.ffzg.hr/bonito2/run.cgi/first_form
Version	3.0
Last update	2013-01-30

Contacts

Marko Tadić	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hnk.ffzg.hr/mt
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics

Distribution

Availability	Available – restricted use	
IPR holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics

	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/
--	----------------	---

Licences

Proprietary		
Restrictions of use	Academic – non-commercial use	
Access medium	Accessible through interface	
Execution location	http://filip.ffzg.hr/bonito2/run.cgi/first_form	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Metadata

Creation date	2013-01-30	
Metadata creators	Marko Tadić	
	Position	Head of the Chair for Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
Metadata language ID	en	
Metadata last date updated	2013-02-04	

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb

		zzl@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	1998-12-01	

Resource documentation

Reports	<p>Marko Tadić. Building the Croatian National Corpus. LREC2002 Proceedings, Las Palmas-Pariz, 2002, Vol. II, 2002, pp 441-446</p> <p>Marko Tadić. Developing the Croatian National Corpus and Beyond. Grzybek, Peter (ed.) Contributions to the Science of Text and Language. Word Length Studies and Related Issues, Kluwer, Dordrecht 2006, pp 295-300</p> <p>Marko Tadić. New version of the Croatian National Corpus. Hlaváčková, Dana; Horák, Aleš; Osolsobě, Klara; Rychlý, Pavel (eds.) After Half a Century of Slavonic Natural Language Processing, Masaryk University, Brno, 2009, pp 199-205</p>
----------------	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	latin
Size	170000000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence

	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation – b pos tagging	
	Segmentation level	Word

3.16. Corpus of Narodne novine

General Information

Short name	NNCorp
Description	The Corpus of Narodne novine (NNCorp) is a specialised corpus of contemporary Croatian standard language written texts published since 1990 in the official journal of the Republic of Croatia "Narodne novine". The corpus is automatically lemmatised and MSD tagged. This is a pseudocorpus, only the query interface using Bonito2 web client is available, while the original texts cannot be distributed. Bonito2 web interface gives opportunity to issue complex queries, thanks to elaborated query language, resulting not only in concordances, but also in word-lists, collocations and other types of distributional data etc. of tokens, lemmas and/or MSDs. The NNCorp v1.0 was entirely included in HNK v2.5.
Identifier	316
Resource type	Corpus
URL	http://filip.ffzg.hr/bonito2/run.cgi/first_form?corpname=nn1990-2005
Version	2.0
Last update	2013-01-30

Contacts

Marko Tadić	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hnk.ffzg.hr/mt
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics

Distribution

Availability	Available – restricted use
---------------------	----------------------------

IPR holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzk@ffzg.hr http://hmk.ffzg.hr/

Licences

Proprietary		
Restrictions of use	Academic – non-commercial use	
Access medium	Accessible through interface	
Execution location	http://filip.ffzg.hr/bonito2/run.cgi/first_form	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzk@ffzg.hr http://hmk.ffzg.hr

Metadata

Creation date	2013-01-30	
Metadata creators	Marko Tadić	
	Position	Head of the Chair for Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
Metadata language ID	en	
Metadata last date updated	2013-02-04	

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG

	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2000-03-01	

Resource documentation

Reports	Marko Tadić. New version of the Croatian National Corpus. Hlaváčková, Dana; Horák, Aleš; Osolsobě, Klara; Rychlý, Pavel (eds.) After Half a Century of Slavonic Natural Language Processing, Masaryk University, Brno, 2009, pp 199-205
----------------	---

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	latin
Size	30000000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	

	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation – b pos tagging	
	Segmentation level	Word

3.17. Croatian n-grams

General Information

Short name	hrNgrams
Description	This resource contains sets of n-grams of different sizes (from 1 to 3) computed from the Croatian National Corpus v2.5. N-grams were computed both from lowercased text and text in original character case. For every size of n above one (i.e. for bigrams and trigrams), n-grams were computed in two ways: taking to account only those appearing within sentence and across sentence boundaries. Regarding the tokenization of the corpus, token is considered to be a continuous sequence of non-whitespace characters. Punctuation markings are treated as separate tokens. Complex punctuations are tokenized as a sequence of simple punctuations. Resource consists of 10 textual files, each computed with different combination of parameters (i.e. n-gram length, character case, sentence boundaries). Each line in the file represents one unique n-gram and its absolute frequency in the corpus, separated by a tabulator. N-grams are ordered according to their frequency, starting from highest to lowest. The n-grams lists were produced using methodology and tools developed by the CESAR Polish partner IPIPAN.
Identifier	317
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://meta-share.ffzg.hr/repository/hrNgrams
Version	1.0

Contacts

Marko Tadić	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics

Distribution

Availability	Available – restricted use	
IPR holder	Marko Tadić	
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr
	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Licences

CC-BY-SA		
Restrictions of use	Attribution	
	Share alike	
Access medium	Downloadable	
Execution location	http://meta-share.ffzg.hr/repository/hrNgrams/	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Metadata

Creation date	2013-01-30	
Metadata creators	Marko Tadić	
	Position	Head of the Chair of Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences

	Department/Institute of Linguistics
Metadata language ID	en
Metadata last date updated	2013-02-04

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzk@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-10-01	

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Syntax
	Linguistic information	Usage – collocations Usage – frequency
Creation	Original source	N-gram list generated using the methodology developed by CESAR Polish partners IPIPAN.
	Creation mode	Mixed
	Creation tools	N-gram generating scripts

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	latin
Size	8681475 entries	
Character encoding	UTF-8	

3.18. Croatian Morphological Lexicon v5.0

General Information

Short name	HML5
Description	The Croatian Morphological Lexicon is an inflectional lexicon generated automatically by Croatian Inflectional Generator from ca 125,000 lemmas yielding over 5,000,000 word forms. It has been a result of the group lead by Marko Tadić on the basis of theoretical background published in 1992 (see Tadić 1994 below). The initial set of lemmas was collected from several existing Croatian mono- and bi-lingual dictionaries, while additional entries were collected via corpus or by means of automatic enlargement of the initial list of lemmas (see Bekavac, Šojat 2005, and Oliver, Tadić 2004 below). The automatically generated output was corrected for known systemic errors, encoded in utf-8 and stored in MulTextEast Lexica format: lemma[TAB]word-form[TAB]MSD. The MSD-tagset is conformant with the MulTextEast v4.0 recommendations for Croatian language. However, some additions exist: in surnames gender is left unspecified (-), additional subclassification of adverbials has been introduced etc. At the moment the Croatian Morphological Lexicon is distributed under CC-BY-NC-SA license.
Identifier	318
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://meta-share.ffzg.hr/repository/hml5
Version	5.0

Contacts

Marko Tadić	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt

Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
---------------------	--

Distribution

Availability	Available – restricted use	
IPR holder	Marko Tadić	
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr
	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Licences

CC-BY-NC-SA		
Restrictions of use	Academic – non-commercial use Attribution Share alike	
Access medium	Downloadable	
Execution location	http://meta-share.ffzg.hr/repository/hml5	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Metadata

Creation date	2013-01-30	
Metadata creators	Marko Tadić	
	Position	Head of the Chair of Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3

		10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
Metadata language ID	en	
Metadata last date updated	2013-02-04	

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	1992-04-01	

Resource documentation

Reports	<p>Marko Tadić. Računalna obradba morfologije hrvatskoga književnoga jezika. PhD Thesis, University of Zagreb, Faculty of Humanities and Social Sciences, 1994.</p> <p>Marko Tadić, Sanja Fulgosi. Building the Croatian Morphological Lexicon. Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages, ACL, Budapest, 2003, pp 41-46</p> <p>Antoni Oliver, Marko Tadić. Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora. LREC2004 Proceedings, Lisbon-Paris, 2004, Vol. IV, pp 1259-1262</p> <p>Bekavac, Božo; Šojat, Krešimir. Lexical acquisition through particular adjectival endings</p>
----------------	---

	<p>for Croatian. Workshop on Computational Modeling of Lexical Acquisition, Split, 2005.</p> <p>Marko Tadić. The Croatian Lemmatization Server. Southern Journal of Linguistics 29 (2005), 1/2, pp 206-217</p> <p>Marko Tadić. Croatian Lemmatization Server. Mila Dimitrova Vulchanova, Svetla Koeva, Iliyana Krapova, Valentin Vulchanov (eds.). Formal Approaches to south Slavic and Balkan Languages, Bulgarian Academy of Sciences, Sofia, 2006, pp 140-146</p>
--	---

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology
	Linguistic information	<p>Lemma</p> <p>Lemma – abbreviations</p> <p>Auxiliary</p> <p>Case</p> <p>Degree</p> <p>Gender</p> <p>Inflection</p> <p>Mood</p> <p>Number</p> <p>Person</p> <p>Tense</p> <p>Other</p> <p>Part of speech</p>
Creation	Original source	Headword lists from different Croatian monolingual dictionaries, Croatian National Corpus, Croatian Web-Corpus
	Creation mode	Mixed
	Creation tools	Croatian Word-Forms Generator

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	latin
Size	5000000 entries	
Character encoding	UTF-8	

3.19. Orwell 1984 Croatian

General Information

Short name	hr1984
Description	The Croatian Orwell 1984 is a Croatian contribution to the MULTEXT-East resources, a multilingual dataset for language engineering research and development. This dataset contains linguistically annotated translations of Orwell's novel 1984 in Bulgarian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Romanian, Serbian, Slovak, Slovene. This corpus adds the Croatian version to the set. The texts in this corpus are lemmatized and MSD-tagged following MTE v4.0 specifications.
Identifier	319
Resource type	Corpus
URL	http://filip.ffzg.hr/bonito2/run.cgi/first_form
Version	4.0
Last update	2013-01-30

Contacts

Marko Tadić	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics

Distribution

Availability	Available – restricted use	
IPR holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr/

Licences

Proprietary	
Restrictions of use	Academic – non-commercial use Attribution
Access medium	Downloadable

Execution location	http://nl.ijs.si/ME/download/licence http://meta-share.ffzg.hr/repository/hr1984	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Metadata

Creation date	2013-01-30	
Metadata creators	Marko Tadić	
	Position	Head of the Chair for Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
Metadata language ID	en	
Metadata last date updated	2013-02-04	

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds
	Funder	European Commission (50%)

		University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	1997-10-01	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	latin
Size	106632 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmaization	
	Segmentation level	Word
	Morphosyntactic annotation – b pos tagging	
	Segmentation level	Word

3.20. Croatian WordNet

General Information

Short name	CroWN
Description	Croatian wordnet (CroWN) is a semantic network of Croatian lexis. CroWN was built on the basis of Princeton WordNet v. 2.0 and synchronized at the end with the Princeton WordNet v. 3.0. Thus, CroWN is completely compatible with PWN 3.0 and, consequently, with all other wordnets mapped to it. It comprises 31,300 literals in 10,040 synsets. Out of all literals, 16,757 (53.54%) of them are nouns, 13,680 (43.7%) verbs,

	857 (2.73%) adjectives and 6 (0.02%) of them are adverbs. As such, CroWN covers 98.87% of synsets from BCS 1, 2 and 3. This resource contains three export files from Croatian wordnet. All three files contain the same data. Two are of the XML type, but formatted according to different schema, and the third is in the JSON format. Collaborators on CroWN development were: Ida Raffaelli, Krešimir Šojat, Daniela Katunar, Matea Srebačić, Vanja Štefanec, Ana Agić, Daša Berović, Lejla Čolić, Marko Tadić, Božo Bekavac, Željko Agić, Igor Marko Gligorić, Ana Ban.
Identifier	320
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Wordnet
URL	http://meta-share.ffzg.hr/repository/CroWN
Version	1.0
Last update	2013-01-30

Contacts

Ida Raffaelli	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb ida.raffaelli@ffzg.hr http://hmk.ffzg.hr/ir
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
Marko Tadić	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics

Distribution

Availability	Available – restricted use	
IPR holder	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr

	http://hmk.ffzg.hr
Availability start date	2013-01-31

Licences

CC-BY-NC-SA	
Restrictions of use	Academic – non-commercial use Attribution Share alike
Access medium	Downloadable
Execution location	http://meta-share.ffzg.hr/repository/CroWN/
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences
	Short name FFZG
	Department name Department/Institute of Linguistics
	Contact Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr

Metadata

Creation date	2013-01-30
Metadata creators	Marko Tadić
	Position Head of the Chair of Algebraic and Computational Linguistics
	Contact Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
	Organization University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
Metadata language ID	en
Metadata last date updated	2013-02-06

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use
---------------------	-----------

	NLP applications
Actual uses	Human use
	NLP applications

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzk@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
	Creation start date 2007-09-01	

Resource documentation

Reports	<p>Bekavac, Božo; Šojat, Krešimir; Tadić, Marko. Zašto nam treba hrvatski WordNet?. In: Granić, Jagoda (ed.) Semantika prirodnog jezika i metajezik semantike, Proceedings of HDPL Annual Conference 2004, Zagreb-Split 2005, pp. 733-743</p> <p>Raffaelli, Ida; Bekavac, Božo; Agić, Željko; Tadić, Marko. Building Croatian WordNet. In: Tanács, Attila; Csendes, Dóra; Vincze, Veronika; Fellbaum, Christianne; Vossen, Piek (eds.) Proceedings of the Fourth Global WordNet Conference 2008, GWC, Szeged 2008, pp. 349-359</p>
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Wordnet	
Lexical conceptual resource encoding	Encoding level	Semantics
	Linguistic	Part of speech

	information	Semantics – cross references Semantics – relations Semantics – relations – antonyms Semantics – relations – hyperonyms Semantics – relations – hyponyms Semantics – relations – meronyms Semantics – relations – synonyms
	Conformance to standards best practices	Word net

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	Latin
Modality	Modality type	Written language
Size	10040 synsets	
Character encoding	UTF-8	

3.21. Croatian Automatic Collocations Dictionary

General Information

Short name	hrACD
Description	The Croatian Automatic Collocations Dictionary has been created by Lexical Computing Ltd. and have been made available to the research community as part of the CESAR project deliverables.
Identifier	321
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://meta-share.ffzg.hr/repository/hrACD

Contacts

Marko Tadić	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3

	10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics

Distribution

Availability	Available – restricted use	
IPR holder	Lexical Computing Ltd.	
	Contact	71, Freshfield Road BN2 0BL Brighton inquiries@sketchengine.co.uk http://www.sketchengine.co.uk/
Availability end date	2012-11-30	

Licences

CC-BY-SA	
Restrictions of use	Attribution Share alike
Access medium	Downloadable
Execution location	http://meta-share.ffzg.hr/repository/hrACD

Metadata

Creation date	2011-11-26	
Metadata creators	Marko Tadić	
	Position	Head of the Chair of Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
Metadata language ID	en	
Metadata last date updated	2013-01-31	

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary
---	-----------------------------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	latin
Size	30000 entries	

3.22. Croatian Weather Dialogue Corpus

General Information

Description	The Croatian Weather Dialogue Corpus (CWEDIC) contains recordings of prepared dialogue questions by multiple male and female native Croatian speakers for use in weather information dialogue system.
Identifier	322
Resource type	Corpus
URL	http://meta-share.ffzg.hr/repository/CWEDIC
Version	1.0

Contacts

Sanda Martinčić-Ipšić	
Position	Assistant Professor
Contact	Radmile Matejčić 2 51000 Rijeka smarti@uniri.hr http://www.inf.uniri.hr
Organization	University of Rijeka Department of Informatics smarti@uniri.hr
Miran Pobar	
Position	Assistant
Contact	Radmile Matejčić 2 51000 Rijeka mpobar@inf.uniri.hr http://www.inf.uniri.hr
Organization	University of Rijeka Department of Informatics

	mpobar@inf.uniri.hr
--	--

Distribution

Availability	Available – restricted use	
IPR holder	Sanda Martinčić-Ipšić	
	Position	Assistant Professor
	Contact	Radmile Matejčić 2 51000 Rijeka smart@uniri.hr
	Organization	University of Rijeka Department of Informatics smart@uniri.hr
	Miran Pobar	
	Position	Assistant
	Contact	Radmile Matejčić 2 51000 Rijeka mpobar@inf.uniri.hr
	Organization	University of Rijeka Department of Informatics mpobar@inf.uniri.hr

Licences

CC-BY-SA	
Restrictions of use	Attribution Share alike
Access medium	Downloadable
Download location	http://meta-share.ffzg.hr/repository/CWEDIC

Metadata

Creation date	2013-02-04	
Metadata creators	Sanda Martinčić-Ipšić	
	Position	Assistant Professor
	Contact	Radmile Matejčić 2 51000 Rijeka smart@uniri.hr http://www.inf.uniri.hr
	Organization	University of Rijeka Department of Informatics

	smarti@uniri.hr
Miran Pobar	
Position	Assistant
Contact	Radmile Matejčić 2 51000 Rijeka mpobar@inf.uniri.hr http://www.inf.uniri.hr
Organization	University of Rijeka Department of Informatics mpobar@inf.uniri.hr
Marko Tadić	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department of Linguistics marko.tadic@ffzg.hr
Metadata language name	English
Metadata language ID	en
Metadata last date updated	2013-02-04

Resource creation

Resource creator	Sanda Martinčić-Ipšić	
	Position	Assistant Professor
	Contact	Radmile Matejčić 2 51000 Rijeka smarti@uniri.hr http://www.inf.uniri.hr
	Organization	University of Rijeka Department of Informatics smarti@uniri.hr

Texts

Media type	text
Linguality type	Monolingual

Languages	Croatian	
	Language ID	hr
	Language script	Latin
	Size	5136 tokens
Modality	Modality type	Spoken language
Size	5135 tokens, 56 minutes	
Character encoding	UTF-8	
Creation	Creation mode	Mixed
Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	Latin
	Size	5135 tokens
Modality	Modality type	Spoken language
Size	5135 tokens, 56 minutes	
Character encoding	UTF-8	
Creation	Creation mode	Mixed

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	Latin
	Size	56 minutes
Modality	Modality type	Spoken language
Audio size	5135 tokens (56 minutes of effective speech in 56 minutes of audio content)	
Audio content	Speech items	Free speech
	Noise level	Low
Setting	Naturalness	Read speech
	Conversational type	Monologue
	Scenario type	Other

	Audience	Large public
	Interactivity	Non interactive
Audio formats	Wave/audio	
	Signal encoding	Other
	Sampling rate	16000
	Quantization	16
	Sign convention	Signed integer
	Number of tracks	1
Creation	Creation mode	Mixed

3.23. CESAR Aligned Wikipedia Headwords List

General Information

Short name	CESAR_WikiHeads
Description	The 762,662 entries of the lexicon are built from the Wikipedia dumps of the six CESAR languages by using article titles and interlingual links to English and the remaining five CESAR languages. In the first phase one lexicon for each CESAR language is built after which those lexicons are merged by grouping together all entries that are connected by interlingual links. If more than one article of a language is connected to a group of articles in other languages (which are actually errors in the structure of the Wikipedias), all article titles are retained, divided by a semicolon. An example of such an entry is "Астеци; Империја Астека". In the final phase category information from the English Wikipedia is added with categories divided by semicolons, and for each non-English entry the number of links to that page in the Wikipedia of the respective language is given.
Identifier	323
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://meta-share.ffzg.hr/repository/CESAR_WikiHeads
Version	1.0

Contacts

Marko Tadić	
Position	Head of the Chair of Algebraic and Computational Linguistics
Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics

Distribution

Availability	Available – restricted use	
IPR holder	Marko Tadić	
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hnk.ffzg.hr
	Nikola Ljubešić	
	Position	Assistant Professor
	Contact	Ivana Lučića 3 10000 Zagreb nljubesi@ffzg.hr http://www.nljubestic.net/
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences
	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hnk.ffzg.hr

Licences

CC-BY-SA	
Restrictions of use	Attribution Share alike
Access medium	Downloadable
Download location	http://meta-share.ffzg.hr/repository/CESAR_WikiHeads

Metadata

Creation date	2013-01-30	
Metadata creators	Marko Tadić	
	Position	Head of the Chair of Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hnk.ffzg.hr/mt
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences

	Department/Institute of Linguistics
Metadata language ID	en
Metadata last date updated	2013-02-04

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzk@ffzg.hr http://hmk.ffzg.hr
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) University of Zagreb, Faculty of Humanities and Social Sciences (50%)
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-12-01	

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology
	Linguistic information	Lemma
Creation	Original source	Headword lists from Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak Wikipedia, aligned with English through interlingual links between Wikipedia articles
	Creation mode	Mixed
	Creation tools	scripts

Texts

Media type	text	
Linguality type	Multilingual	
Languages	Bulgarian	
	Language ID	bg
	Language script	cyrillic
	Croatian	
	Language ID	hr
	Language script	latin
	English	
	Language ID	en
	Language script	latin
	Hungarian	
	Language ID	hu
	Language script	latin
	Polish	
	Language ID	pl
	Language script	latin
	Serbian	
	Language ID	sr
	Language script	cyrillic
	Slovakian	
	Language ID	sk
	Language script	latin
Size	762662 entries	
Character encoding	UTF-8	

3.24. Croatian and Slovene NERC models for Stanford NERC

General Information

--	--

Short name	hrStanfordNERC, slStanfordNERC
Description	Stanford NER model for named entity recognition and classification (NERC) in Croatian texts is built by using the Stanford Named Entity Recognizer tool (URL http://nlp.stanford.edu/software/CRF-NER.shtml) based on Conditional Random Fields (CRF). The model was trained on a portion of texts crawled from the Vjesnik news portal and manually annotated for ENAMEX TYPE={LOCATION, ORGANIZATION, PERSON}. The manually tagged portion of the text consists of 200.006 tokens in 7.358 sentences, containing 5.966 person tokens, 6.897 organization tokens and 4.784 location tokens. Tokens and named entity tags were used as features in the training procedure. Stanford NER model for named entity recognition and classification (NERC) in Slovene texts is built by using the Stanford Named Entity Recognizer tool (URL http://nlp.stanford.edu/software/CRF-NER.shtml) based on Conditional Random Fields (CRF). The model was trained on a portion of texts selected from the SSJ-500k corpus of Slovene (URL http://www.slovenscina.eu/tehnologije/ucni-korpus) and manually annotated for ENAMEX TYPE={LOCATION, ORGANIZATION, PERSON, MISC}. The manually tagged portion of the text consists of 216.011 tokens in 9.663 sentences, containing 4.204 person tokens, 2.526 organization tokens, 2.421 location tokens and 1.143 miscellaneous tokens. Manually assigned POS/MSD tags and lemmas for the 216.011 tokens were extracted from the corpus and used as features in the training procedure.
Identifier	324
Resource type	Language description
Language description type	Other
URL	http://meta-share.ffzg.hr/repository/StanfordNERC
Version	1.0
Last update	2013-01-30

Contacts

Božo Bekavac	
Position	Assistant Professor
Contact	Ivana Lučića 3 10000 Zagreb bbekavac@ffzg.hr http://hmk.ffzg.hr/bb
Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics

Distribution

Availability	Available – restricted use	
IPR holder	Božo Bekavac	
	Position	Assistant Professor
	Contact	Ivana Lučića 3 10000 Zagreb

		bbekavac@ffzg.hr http://hmk.ffzg.hr/bb
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics

Licences

ApacheLicence_2.0		
Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://meta-share.ffzg.hr/repository/StanfordNERC	
Signatories	University of Zagreb, Faculty of Humanities and Social Sciences	
	Short name	FFZG
	Department name	Department/Institute of Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hmk.ffzg.hr
User nature	Academic	

Metadata

Creation date	2013-02-05	
Metadata creators	Marko Tadić	
	Position	Head of the Chair of Algebraic and Computational Linguistics
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hmk.ffzg.hr/mt
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
Metadata language ID	en	
Metadata last date updated	2013-02-05	

Usage

Access tool	Stanford NERC system http://nlp.stanford.edu/software/CRF-NER.shtml
--------------------	---

Resource creation

--	--

Resource creator	Božo Bekavac	
	Position	Assistant Professor
	Contact	Ivana Lučića 3 10000 Zagreb bbekavac@ffzg.hr http://hmk.ffzg.hr/bb
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department/Institute of Linguistics
	Željko Agić	
	Position	Senior Assistant
	Contact	Ivana Lučića 3 10000 Zagreb zagic@ffzg.hr http://www.ffzg.unizg.hr/infoz/hr/index.php/lanovi-odsjeaka/171-eljko-agi
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences Department of Information and Communication Sciences
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net/
	Funding type	EU funds National funds Own funds
	Funder	European Commission University of Zagreb, Faculty of Humanities and Social Sciences
	Country	EU Croatia
	Start date	2011-02-01
	End date	2013-01-31

Language description

Language description type	Other	
Texts	Media type	text
	Linguality type	Multilingual
	Multilinguality type	Other
	Languages	Croatian
	Language ID	hr

		Slovenian	
		Language ID	sl

Texts

Media type	text		
Linguality type	Multilingual		
Multilinguality type	Other		
Languages	Croatian		
	Language ID	hr	
	Slovenian		
	Language ID	sl	

3.25. Coral Corpus Aligner

General Information

Short name	Coral
Description	Coral (CORpus ALigner) is a tool for easy bilingual parallel corpora alignment. It allows both automatic and manual alignment. The main features of Coral are: (1) Automatic segmentation of texts into sentences, (2) manual sentencesegmentation editing, (3) automatic parallel text alignment using either theGale-Church alignment method or a naïve one-on-one alignment approach, (4) anextremely easy to use manual sentence alignment user interface, (5) exportsalignment results into a standard TMX file, (6) runs on all operating systemsthat can run the Java Virtual Machine, (7) easy installation (a .zip file issimply unpacked onto the user's machine).
Identifier	325
Resource type	Tool/service
Tool/service type	Tool
URL	http://takelab.fer.hr/coral
Version	1.0

Contacts

Jan Šnajder	
Position	Assist. prof.
Contact	Unska 3 10000 Zagreb jan.snajder@fer.hr http://www.fer.unizg.hr/jan.snajder

Organization	Univ. of Zagreb, Faculty of Electrical Engineering and Computing, Text Analysis and Knowledge... jan.snajder@fer.hr
---------------------	--

Distribution

Availability	Available – restricted use	
IPR holder	Jan Šnajder	
	Contact	Unska 3 10000 Zagreb jan.snajder@fer.hr http://www.fer.unizg.hr/jan.snajder
	Bojana Dalbelo Bašić	
	Contact	Unska 3 10000 Zagreb bojana.dalbelo@fer.hr http://www.fer.unizg.hr/bojana.dalbelo-basic
	Marin Akšamović	
	Contact	marin.aksamovic@gmail.com
	Renato Dragišić	
	Contact	renato.dragisic@gmail.com
	Đive Jakobušić	
	Contact	dive.jakobusic@gmail.com
	Marin Japiec	
	Contact	marin.japiec+coral@gmail.com
	Vjekoslav Osmann	
	Contact	vosmann@gmail.com
	Željko Rumenjak	
	Contact	zeljko.rumenjak@gmail.com
	Ivan Šolta	
	Contact	ivan.solta@gmail.com
	Igor Šoš	
	Contact	igosos@gmail.com
	Univ. of Zagreb, Faculty of Electrical Engineering and Computing, Text Analysis and Knowledge...	
	Contact	Unska 3 10000 Zagreb takelab@fer.hr http://takelab.fer.hr

Licences

Restrictions of use	Inform licensor	
Access medium	Downloadable	
Download location	http://takefab.fer.hr/coral	
Distribution rights holder	Univ. of Zagreb, Faculty of Electrical Engineering and Computing, Text Analysis and Knowledge...	
	Contact	Unska 3 10000 Zagreb takefab@fer.hr http://takefab.fer.hr

Metadata

Creation date	2013-01-24	
Metadata creators	Jan Šnajder	
	Position	assistant professor
	Contact	Unska 3 10000 Zagreb jan.snajder@fer.hr http://www.fer.unizg.hr/jan.snajder
	Organization	Univ. of Zagreb, Faculty of Electrical Engineering and Computing, Text Analysis and Knowledge... jan.snajder@fer.hr
Metadata language ID	en	
Metadata last date updated	2013-02-04	

Resource creation

Resource creator	Jan Šnajder	
	Contact	Unska 3 10000 Zagreb jan.snajder@fer.hr http://www.fer.unizg.hr/jan.snajder
Funding projects	Otkrivanje znanja u tekstnim podacima / Knowledge discovery in textual data	
	Project short name	MZOS-036-1300646-1986
	Funding type	National funds
	Funder	Ministry of Science, Education and Sports, Republic of Croatia

	Start date	2007-01-01
	End date	2013-06-30
Creation start date	2011-10-01	

Resource documentation

Reports	
----------------	--

Tool/service

Tool/service type	Tool	
Language dependent	False	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	OS-independent	
Tool/service creation	Implementation language	Java

3.26. Croatian Sentiment Lexicon

General Information

Short name	CroSentiLex
Description	CroSentiLex is a sentiment lexicon for Croatian. CroSentiLex consists of two files (crosentilex-positives.txt and crosentilex-negatives.txt), each containing 37K Croatian lemmas ranked by positivity and negativity, respectively, with the corresponding PageRank scores. The rankings were created automatically based on small positive and negative seed sets and co-occurrence frequencies, using the PageRank algorithm. In addition to the automatically extracted lexicon, human (gold-standard) sentiment annotations for 1200 Croatian lemmas are provided in sentsentiment-annotations.txt.
Identifier	326
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://takefab.fer.hr/sentilex
Version	1.0

Contacts

Goran Glavaš

Position	PhD student
Contact	Unska 3 10000 Zagreb goran.glavas@fer.hr http://www.fer.unizg.hr/goran.glavas
Organization	Univ. of Zagreb, Faculty of Electrical Engineering and Computing, Text Analysis and Knowledge... goran.glavas@fer.hr

Distribution

Availability	Available – restricted use	
IPR holder	Jan Šnajder	
	Contact	Unska 3 10000 Zagreb jan.snajder@fer.hr http://www.fer.unizg.hr/jan.snajder
	Bojana Dalbelo Bašić	
	Contact	Unska 3 10000 Zagreb bojana.dalbelo@fer.hr http://www.fer.unizg.hr/bojana.dalbelo-basic
	Goran Glavaš	
	Contact	Unska 3 10000 Zagreb goran.glavas@fer.hr http://www.fer.unizg.hr/goran.glavas
	Univ. of Zagreb, Faculty of Electrical Engineering and Computing, Text Analysis and Knowledge...	
	Contact	Unska 3 10000 Zagreb takefab@fer.hr http://takefab.fer.hr

Licences

CC-BY-NC-SA	
Restrictions of use	Academic – non-commercial use
Access medium	Downloadable
Download location	http://takefab.fer.hr/sentilex
Distribution rights	Univ. of Zagreb, Faculty of Electrical Engineering and Computing, Text Analysis

holder	and Knowledge...	
	Contact	Unska 3 10000 Zagreb takefab@fer.hr http://takefab.fer.hr

Metadata

Creation date	2013-01-24	
Metadata creators	Goran Glavaš	
	Position	PhD student and research assistant
	Contact	Unska 3 10000 Zagreb goran.glavas@fer.hr http://www.fer.unizg.hr/goran.glavas
	Organization	Univ. of Zagreb, Faculty of Electrical Engineering and Computing, Text Analysis and Knowledge... goran.glavas@fer.hr
Metadata language ID	en	
Metadata last date updated	2013-02-04	

Resource creation

Resource creator	Goran Glavaš	
	Contact	Unska 3 10000 Zagreb goran.glavas@fer.hr http://www.fer.unizg.hr/goran.glavas
Funding projects	Otkrivanje znanja u tekstnim podatcima / Knowledge discovery in textual data	
	Project short name	MZOS-036-1300646-1986
	Funding type	National funds
	Funder	Ministry of Science, Education and Sports, Republic of Croatia
	Start date	2007-01-01
	End date	2013-06-30
Creation start date	2011-10-01	

Resource documentation

--	--

Reports	Glavaš, G., Šnajder, J., Dalbelo Bašić, B. Semi-Supervised Acquisition of Croatian Sentiment Lexicon. In Proceedings of 15th International Conference on Text, Speech and Dialogue, TSD 2012, Brno, Czech Republic, September 2012., pp. 166-173
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Semantics
	Linguistic information	Other
Creation	Creation mode	Mixed

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	latin
Size	76778 entries	
Character encoding	UTF-8	

4. IPIAN resources

4.1. Polish Sejm Corpus

General Information

Short name	PSC
Description	The Polish Sejm Corpus contains annotated utterances of Polish Sejm members from terms of office 1-6 (years 1991-2011). Corpus files contain information about text segmentation (paragraphs, sentences, tokens), disambiguated morphosyntactic description (lemma, POS tag, MSD tag), syntactic description (syntactic words and groups) and named entities (person names, locations, organization). The data is a valuable source of linguistic information, being a large (100 M segments) collection of quasi-spoken content and making the basis of the audio/video recording of sessions, started in 2011 and planned to be consecutively appended to the corpus.
Identifier	401
Resource type	Corpus
URL	http://clip.ipipan.waw.pl/PSC

Version	1.0
Last update	2011-11-12

Contacts

Maciej Ogrodniczuk	
Position	Assistant Professor
Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

BSD-style		
Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://clip.ipipan.waw.pl/PSC	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2011-10-17	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw

		maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2011-11-26	

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
Size	2.6 gb, 114000000 tokens	
Annotation	Other	
	Annotation standoff	False
	Segmentation level	Paragraph
	Format	text/xml

Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation mode details	transcripts of one session day represented as a set of files: header.xml – header information about the individual session days, text_structure.xml – session structure and individual utterances, ann_segmentation.xml.gz – compressed sentence-level and token-level segmentation, ann_morphosyntax.xml.gz – disambiguated morphosyntactic description (lemma, POS tag and MSD tag), ann_words.xml.gz – syntactic words, ann_groups.xml.gz – syntactic groups, ann_named.xml.gz – named entities
Annotation tool	scripts developed internally
Start date	2011-03-01
End date	2011-11-26
Other	
Annotation standoff	False
Segmentation level	Paragraph
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation mode details	representation of the session structure (taken over from the transcripts) in <div> elements
Annotation tool	scripts developed internally
Start date	2011-03-01
End date	2011-11-26
Segmentation	
Annotation standoff	False
Segmentation level	Utterance
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Automatic

Annotation mode details	segmentation into utterances taken over from the transcripts; each utterance is marked with the speaker identifier (resolved in the transcript header)
Annotation tool	scripts developed internally
Start date	2011-03-01
End date	2011-11-26
Segmentation	
Annotation standoff	True
Segmentation level	Sentence
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation mode details	individual utterances split into sentences
Annotation tool	Pantera
Start date	2011-03-01
End date	2011-11-26
Segmentation	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation mode details	individual sentences split into tokens (word-like segments – see documentation of Morfeusz SGJP for details)
Annotation tool	Morfeusz SGJP
Start date	2011-03-01
End date	2011-11-26
Lemmatization	

Annotation standoff	True
Segmentation level	Word
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation mode details	lemma variants (all available interpretations) output by Morfeusz, then disambiguated by Pantera tagger
Annotation tool	Morfeusz SGJP
Start date	2011-03-01
End date	2011-11-26
Morphosyntactic annotation - below POS tagging	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation mode details	MSD tag variants (all available morphosyntactic interpretations) output by Morfeusz, then disambiguated by Pantera tagger
Annotation tool	Morfeusz SGJP
Start date	2011-03-01
End date	2011-11-26
Morphosyntactic annotation – POS tagging	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best	TEI

practices	
Annotation mode	Automatic
Annotation mode details	POS tag (CTAG) variants (all available interpretations) output by Morfeusz, then disambiguated by Pantera tagger
Annotation tool	Pantera
Start date	2011-03-01
End date	2011-11-26
Structural annotation	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation mode details	syntactic words (word-like compounds) detected by Spejd with NKJP shallow parsing grammar; see NKJP documentation for details
Annotation tool	Spejd
Start date	2011-03-01
End date	2011-11-26
Syntactic annotation – shallow parsing	
Annotation standoff	True
Segmentation level	Word group
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation mode details	syntactic groups (phrase-like constructs) detected by Spejd with NKJP shallow parsing grammar; see NKJP documentation for details
Annotation tool	Spejd
Start date	2011-03-01
End date	2011-11-26

	Semantic annotation – named entities	
	Annotation standoff	True
	Segmentation level	Word group
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Automatic
	Annotation mode details	named entities (person names, organizations, locations compatible with NKJP hierarchy) detected by Nerf
	Annotation tool	Nerf
	Start date	2011-03-01
	End date	2011-11-26
Text classification	Text type	quasi-spoken
	Register	formal
Creation	Original source	Sprawozdanie Stenograficzne. Kancelaria Sejmu Rzeczypospolitej Polskiej, ul. Wiejska 4/6/8, 00-902, Warszawa, Poland. Wydawnictwo Sejmowe, 1991-2011. ISSN 08672768. http://www.sejm.gov.pl
	Creation mode	Automatic
	Creation mode details	Texts from terms 1-4 covered from HTML files, terms 5-6 converted from XML files delivered by Sejm. Audio and video sample from day 3 of sitting 89, term 6 added as example of multimodal content.
	Creation tools	Morfeusz SGJP, a tokenizer, morphological analyzer and lemmatizer for Polish Pantera, a Brill tagger for Polish Spejd, a shallow parser of Polish Nerf, a named entity recognizer for Polish

4.2. PoliMorf Inflectional Dictionary

General Information

Short name	PoliMorf
Description	The new morphological dictionary for Polish resulting from the standardization, merger and manual correction of Morfeusz SGJP and Morfologik. Morfeusz SGJP is a morphological analyser for Polish whose inflectional data (dictionary) comes from SGJP — Grammatical Dictionary of Polish. SGJP is the result of several years of work of an informal group lead by Prof. Saloni. The work started in the 1980s by digitising the list of headwords of the 11-volume Doroszewski's dictionary of Polish (1958–1969). The grammatical description in SGJP is based on new concepts proposed in the 2nd half of the 20th century with many

	<p>detailed solutions proposed by the members of the team (Tokarski, Gruszczyński, Saloni). PoliMorf uses data from the second edition of SGJP. 244,341 lexemes correspond to 4,223,981 word forms (counting syncretic forms of the same lexeme as one unit). Inflection in SGJP is represented with inflectional patterns, which describe forms in terms of a stem common to all forms and endings differentiating the forms. Morfologik is an open-source morphological dictionary of Polish. It contains 216,992 lexemes and 3,475,809 word forms. The dictionary was created by enriching the Polish ispell/hunspell dictionary with morphological information, which was possible thanks to the structure of the original dictionary that retained important grammatical distinctions. The process of conversion relied on a series of scripts, and the resulting dictionary was later augmented with manually entered information. Unfortunately, the original source dictionary did not contain sufficient structure to allow reliable detection of some information, such as the exact subgender of the masculine for substantives. This information was added manually and using heuristic methods, however its reliability is low. Considering the fact that the substantives are about one third of the dictionary content (and almost half of them are masculine), this limitation is severe. The tagset of the dictionary is inspired by the IPI PAN Tagset. However, Morfologik diverges from that tagset and from Morfeusz, as it never splits orthographic (“space-to-space”) words into smaller dictionary words (i.e. so-called agglutination is not considered). Moreover, due to the lack of information in the ispell dictionary, some forms are not completely annotated, and are marked as irregular. There is, however, some additional mark up added to reflexive verbs, which is not present in the original IPI PAN Tagset. This was introduced for the purposes of the grammar checker LanguageTool that used the dictionary extensively.</p>
Identifier	402
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://zil.ipipan.waw.pl/PoliMorf
Version	0.5

Contacts

Maciej Ogrodniczuk	
Position	Assistant Professor
Contact	<p>Jana Kazimierza 5 01-248 Warsaw</p> <p>maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk</p>
Organization	<p>Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl</p>

Distribution

Availability	Available – unrestricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN

	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

BSD-style		
Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/PoliMorf?action=AttachFile&do=get&target=PoliMorf-0.5.tab.gz	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2011-11-25	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2011-11-26	

Resource creation

Resource creator	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2011-02-01	

Resource documentation

Reports	<p>Marcin Milkowski. Developing an open-source, rule-based proofreading tool. In: Software: Practice & Experience 40 (7), pp. 543-566. http://doi.wiley.com/10.1002/spe.971</p> <p>Zygmunt Saloni, Włodzimierz Gruszczyński, Marcin Woliński and Robert Wołosz. Słownik gramatyczny języka polskiego. Wiedza Powszechna. Warszawa 2007.</p>
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology
	Linguistic information	<p>Lemma</p> <p>Aspect</p> <p>Auxiliary</p> <p>Case</p> <p>Degree</p>

		Gender Inflection Mood Number Person Tense Part of speech
Creation	Original source	Morfęusz SGJP Morfologik
	Creation mode	Mixed
	Creation tools	Kuźnia

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	latin
Size	36.3 mb, 6382227 entries	
Character encoding	UTF-8	

4.3. Polish WordNet

General Information

Short name	plWordNet
Description	The Polish WordNet is a network of lexical-semantic relations, an electronic thesaurus with a structure modelled on that of the Princeton WordNet and those constructed in the EuroWordNet project. Polish WordNet describes meaning of a lexical unit by placing it within a network of semantic relations, such as hypernymy, meronymy, antonymy etc. To reduce the cost of the project, Polish WordNet has been built semi-automatically. Lexical relations were automatically recognized in large corpora of Polish and suggested to linguists/lexicographers via a graphical interface. Nowadays Polish WordNet is one of the biggest wordnets in the world; it comprises 160000 lexical units in 116000 synsets. The current version of the resource introduces 54000 relations between synsets of plWordNet and Princeton WordNet.
Identifier	403
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Wordnet

URL	http://plwordnet.pwr.wroc.pl/wordnet
Version	1.7
Last update	2012-07-27

Contacts

Maciej Piasecki	
Contact	Wybrzeże Wyspiańskiego 27 50-370 Wrocław maciej.piasecki@pwr.wroc.pl http://www.iis.pwr.wroc.pl/~piasecki/
Organization	Wrocław University of Technology Institute of Informatics, Division of Artificial Intelligence maciej.piasecki@pwr.wroc.pl

Distribution

Availability	Available – restricted use	
IPR holder	Wrocław University of Technology	
	Contact	Wybrzeże Wyspiańskiego 27 50-370 Wrocław maciej.piasecki@pwr.wroc.pl http://nlp.pwr.wroc.pl
Availability start date	2011-11-22	

Licences

Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://nlp.pwr.wroc.pl/plwordnet/download/?lang=eng	
Fee	free of charge	
Signatories	Maciej Piasecki	
	Contact	Wybrzeże Wyspiańskiego 27 50-370 Wrocław maciej.piasecki@pwr.wroc.pl http://www.iis.pwr.wroc.pl/~piasecki/
Distribution rights holder	Wrocław University of Technology	
	Contact	Wybrzeże Wyspiańskiego 27 50-370 Wrocław maciej.piasecki@pwr.wroc.pl

	http://www.iis.pwr.wroc.pl/~piasecki/
--	---

Metadata

Creation date	2011-11-22	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2011-11-26	

Usage

Foreseen use	Human use NLP applications
NLP-specific use	Document classification Semantic role labelling

Resource creation

Resource creator	Maciej Piasecki	
	Contact	Wybrzeże Wyspiańskiego 27 50-370 Wrocław maciej.piasecki@pwr.wroc.pl http://www.iis.pwr.wroc.pl/~piasecki/
Funding projects	NEKST — Adaptive system supporting solving problems on the basis of content analysis of electronic documents	
	Project ID	POIG.01.01.02-14-013/09
	Funding type	EU funds National funds
	Country	Poland
	Start date	2009-04-01
	End date	2014-02-10
	Automatic methods of constructing a semantic network of Polish lexemes for natural language processing	

	Project ID	3 T11C 018 29
	Funding type	National funds
	Country	Poland
	Start date	2005-10-31
	End date	2008-10-30
	Construction of lexical resources with the help of recognition of semantic relations in text corpora on the basis of morpho-syntactic and semantic data	
	Project ID	N N516 068637
	Funding type	National funds
	Country	Poland
	Start date	2009-10-31
	End date	2012-10-30
	SyNaT — Research Task: “Construction of an open, repository hosting and communication platform for the network knowledge resources fro science, education and open knowledge society”	
	Project ID	SP/I/1/77065/10
	Funding type	National funds
	Country	Poland
	Start date	2010-08-16
	End date	2013-08-16
Creation start date	2005-10-31	

Resource documentation

Reports	<p>Maziarz M., Piasecki M., Szpakowicz S. Approaching plWordNet 2.0. Proceedings of the 6th Global Wordnet Conference, Matsue, 9-13th January, 2012, Japan.</p> <p>Piasecki, Maciej, Szpakowicz, Stanisław, Bartosz Broda. A Wordnet from the Ground Up. Wrocław : Oficyna Wydawnicza Politechniki Wrocławskiej, 2009.</p>
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Wordnet	
Lexical conceptual resource encoding	Encoding level	Semantics
	Linguistic information	Part of speech Semantics – cross references Semantics – relations Semantics – relations – antonyms Semantics – relations – hyperonyms Semantics – relations – hyponyms Semantics – relations – meronyms

		Semantics – relations – synonyms
	Conformance to standards best practices	Word net

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	Latin
Modality	Modality type	Written language
Size	116000 synsets, 106500 lemmata (strings in Princeton WordNet), 160000 lexical units (i.e., lemma-sense pairs), 54000 relations between Polish and English synsets	

4.4. Polish Named Entity Recognition Tool

General Information

Short name	Nerf
Description	Nerf is a statistical tool for named entity (NE) recognition based on linear-chain conditional random fields. The tool has been constructed as a part of the National Corpus of Polish project and then reimplemented and improved within the CESAR project. It has been adapted to recognize tree-like structures of NEs (i.e., with recursively embedded NEs) by using the joined label tagging method. The tool supports a collection of Polish NE-related dictionaries which significantly improve the quality of NE recognition.
Identifier	404
Resource type	Tool/service
Tool/service type	Tool
URL	http://zil.ipipan.waw.pl/Nerf
Version	0.2
Last update	2011-10-11

Contacts

Jakub Waszczuk	
Contact	Jana Kazimierza 5 01-248 Warsaw waszczuk.kuba@gmail.com

Distribution

Availability	Available – restricted use
--------------	----------------------------

Licences

BSD	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://clip.ipipan.waw.pl/Nerf
Fee	free of charge

Metadata

Creation date	2011-11-24	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Jakub Waszczuk	
	Contact	Jana Kazimierza 5 01-248 Warsaw waszczuk.kuba@gmail.com
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2011-11-24	

Usage

Foreseen use	NLP applications	
NLP-specific use	Named entity recognition	
Actual uses	NLP applications	
	NLP-specific use	Named entity recognition
	Reports	Jakub Waszczuk, Katarzyna Głowińska, Agata Savary and Adam Przepiórkowski. 2010. Tools and Methodologies for Annotating Syntax and Named Entities in the National Corpus of Polish. In Proceedings of the International Multiconference on Computer Science and Information

	Technology (IMCSIT 2010): Computational Linguistics – Applications (CLA’10), pages 531–539, Wisła, Poland. PTI. Agata Savary, Jakub Waszczuk and Adam Przepiórkowski. 2010. Towards the Annotation of Named Entities in the National Corpus of Polish. In Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta. ELRA.	
Usage project	National Corpus of Polish	
	Project short name	NKJP
	URL	http://www.nkjp.pl
	Funding type	National funds
	Funder	The Polish Ministry of Science and Higher Education (100%)
	Country	Poland
	Start date	2007-12-13
	End date	2011-06-12
Actual use details	Recognition of Named Entities in the National Corpus of Polish. Tool trained on the manually annotated million-word subcorpus has been used to annotate the entire NKJP corpus.	
NLP applications		
NLP-specific use	Named entity recognition	
Reports	Ogrodniczuk M., Przepiórkowski A. Polish Language Processing Chains for Multilingual Information Systems. G. Bouma et al. (ed.): NLDB 2012, LNCS 7337, pp. 152–157. Springer, Heidelberg.	
Usage project	Applied Technology for Language-Aided CMS	
	Project short name	ATLAS
	URL	http://www.atlasproject.eu
	Funding type	EU funds National funds
	Funder	European Commission (50%) The Polish Ministry of Science and Higher Education (50%)
	Country	Poland
	Start date	2010-03-01
	End date	2013-02-28
Actual use details	Recognition of Named Entities for the UIMA Language Processing Chain in ATLAS CMS.	

NLP applications		
NLP-specific use	Parsing	
Reports	Ogrodniczuk M., Kopeć M. Rule-based coreference resolution module for Polish. In Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), pp. 191–200. Faro, Portugal.	
Usage project	Computer-based methods for coreference resolution in Polish texts	
	Project short name	CORE
	URL	http://zil.ipipan.waw.pl/CORE
	Funding type	National funds
	Funder	National Science Centre (100%)
	Country	Poland
	Start date	2011-04-18
	End date	2014-04-17
Actual use details	Named entity-based mention detection for the Polish coreference resolution module.	

Resource creation

Resource creator	Jakub Waszczuk	
	Contact	Jana Kazimierza 5 01-248 Warsaw waszczuk.kuba@gmail.com
	Michał Lenart	
	Position	Programmer
	Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@gmail.com http://zil.ipipan.waw.pl/MichalLenart
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%)

		Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
	National Corpus of Polish	
	Project short name	NKJP
	URL	http://nkjp.pl/
	Funding type	National funds
	Funder	Polish Ministry of Science and Higher Education
	Country	Poland
	Start date	2007-01-01
	End date	2011-06-30
Creation start date	2010-03-01	

Resource documentation

Reports	Jakub Waszczuk, Katarzyna Głowińska, Agata Savary, Adam Przepiórkowski, "Tools and methodologies for annotating syntax and named entities in the National Corpus of Polish", Proceedings of the 2010 International Multiconference on Computer Science and Information Technology (IMCSIT).
Tool documentation type	Help functions

Tool/service

Tool/service type	Tool	
Language dependent	False	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	OS-independent	
Required software	The Glasgow Haskell Compiler (version 7.0.4 or higher)	
Tool/service evaluation	Evaluated	True
	Level	Diagnostic
	Type	Black box
	Criteria	Intrinsic

	Measure	Automatic		
	Reports	Results of Nerf evaluation on the National Corpus of Polish has been described in: Agata Savary and Jakub Waszczuk. Narzędzia do anotacji jednostek nazewniczych. In Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors, Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish], pages 225–252. Wydawnictwo Naukowe PWN, Warsaw, 2012.		
	Details	Cross validation of the Nerf tool on the NKJP corpus, with respect to the NKJP named entities hierarchy, yielded F-measure of 79%.		
	Evaluators	Jakub Waszczuk <table><tr><td>Contact</td><td>Jana Kazimierza 5 01-248 Warsaw waszczuk.kuba@gmail.com</td></tr></table>		Contact
Contact	Jana Kazimierza 5 01-248 Warsaw waszczuk.kuba@gmail.com			
Tool/service creation	Implementation language	Haskell		
	Formalism	conditional random fields joined label tagging		

4.5. 1 million subcorpus of National Corpus of Polish

General Information

Short name	1MNKJP
Description	The National Corpus of Polish (PL: Narodowy Korpus Języka Polskiego, NKJP) is a shared initiative of four institutions: Institute of Computer Science at the Polish Academy of Sciences (coordinator), Institute of Polish Language at the Polish Academy of Sciences, Polish Scientific Publishers PWN, and the Department of Computational and Corpus Linguistics at the University of Łódź. It has been registered as a research-development project of the Ministry of Science and Higher Education. The list of sources for the corpus contains classic literature, daily newspapers, specialist periodicals and journals, transcripts of conversations, and a variety of short-lived and internet texts. The resources represent wide diversity with respect to the subject and genre. The spoken part covers both male and female speakers, in various age groups, coming from various regions in Poland. The 1-million subcorpus of NKJP has been manually annotated.
Identifier	405
Resource type	Corpus
URL	http://www.nkjp.pl
Version	1.0

Contacts

Adam Przepiórkowski	
Position	Professor, Head of the Linguistic Engineering Group

Contact	Jana Kazimierza 5 01-248 Warsaw adam.przepiorkowski@ipipan.waw.pl http://zil.ipipan.waw.pl/AdamPrzepiorkowski
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – unrestricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

GPL		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://clip.ipipan.waw.pl/LRT?action=AttachFile&do=view&target=NKJP-PodkorpusMilionowy-1.1.tgz	
Fee	free of charge	
Signatories	Adam Przepiórkowski	
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2011-11-12	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw

	maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Łukasz Degórski
Position	Assistant
Contact	Jana Kazimierza 5 01-248 Warsaw ldegorski@ipipan.waw.pl http://zil.ipipan.waw.pl/LukaszDegorski
Source	CESAR
Metadata language ID	en
Metadata last date updated	2011-11-26

Texts

Media type	text
Linguality type	Monolingual
Languages	Polish
	Language ID PL
Size	165.0 mb, 1003956 words
Annotation	Segmentation
	Annotation standoff True
	Segmentation level Paragraph
	Format text/xml
	Conformance to standards best practices TEI
	Annotation mode details inherited from source corpus (no need to generate new segmentation when sampling)
	Start date 2009-06-29
	End date 2010-05-21
	Segmentation
	Annotation standoff True
	Segmentation level Sentence
	Format text/xml

Conformance to standards best practices	TEI
Annotation mode	Manual
Annotation mode details	annotated manually using Anotatornia tool
Annotation tool	Anotatornia
Start date	2009-06-29
End date	2010-09-30
Segmentation	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode	Mixed
Annotation mode details	paragraphs split into tokens (word-like segments – see documentation of Morfeusz SGJP for details); segmentation revised by annotators using Anotatornia
Annotation tool	Morfeusz SGJP (automatic), Anotatornia (manual)
Start date	2009-06-29
End date	2010-09-30
Lemmatization	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Mixed
Annotation mode details	lemma variants (all available interpretations) output by Morfeusz, then manually disambiguated using Anotatornia

Annotation tool	Morfeusz SGJP (automatic), Anotatornia (manual)
Start date	2009-06-29
End date	2010-09-30
Morphosyntactic annotation - below POS tagging	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode	Mixed
Annotation mode details	MSD tag variants (all available morphosyntactic interpretations) output by Morfeusz, then manually disambiguated using Anotatornia
Annotation tool	Morfeusz SGJP (automatic), Anotatornia (manual)
Start date	2009-06-29
End date	2010-09-30
Morphosyntactic annotation – POS tagging	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode	Mixed
Annotation mode details	POS tag (CTAG) variants (all available interpretations) output by Morfeusz, then manually disambiguated using Anotatornia
Annotation tool	Morfeusz SGJP (automatic), Anotatornia (manual)
Start date	2009-06-29
End date	2010-09-30
Syntactic annotation – shallow parsing	
Annotation standoff	True

Segmentation level	Word group
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Mixed
Annotation mode details	syntactic words (word-like compounds) and groups output by Spejd, then manually disambiguated using TrEd
Annotation tool	Spejd (automatic), TrEd (manual)
Start date	2010-01-01
End date	2010-09-30
Semantic annotation – named entities	
Annotation standoff	True
Segmentation level	Word group
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Mixed
Annotation mode details	named entities output by Nerf, then manually disambiguated using TrEd
Annotation tool	Nerf (automatic), TrEd (manual)
Start date	2010-01-01
End date	2010-12-31
Semantic annotation – word senses	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Mixed

	Annotation mode details	word sense information output automatically, then manually disambiguated using Anotatornia
	Annotation tool	WSDDE (automatic), Anotatornia (manual)
	Start date	2009-06-29
	End date	2010-12-31
Creation	Original source	the IPI PAN corpus the PELCRA corpus the PWN corpus text collected by IJP PAN, PELCRA and PWN specifically for NKJP
	Creation mode	Mixed
	Creation mode details	Texts from the NKJP have been sampled automatically; samples have been revised manually. Linguistic annotation on all levels has been added manually (possibly basing on some automatic annotation).
	Creation tools	Morfeusz SGJP Spejd Nerf Anotatornia various shell scripts

4.6. Polish Named Entity Gazetteer

General Information

Short name	PNEG
Description	The Polish Named Entity Gazetteer is an electronic lexicon containing partly inflected entries of Polish (and some foreign) proper names and named entity components (forenames and surnames, geographical names, organizational names, relational adjectives and inhabitant names stemming from country names as well as named entity triggers – months, days, positions, etc.). The resource was used for the automatic pre-annotation of the National Corpus of Polish (NKJP) on the level of named entities. The resource is available in: (i) a textual format compliant with the Sprout text processing platform, (ii) an LMF-compliant format. It contains about 45,000 lemmas and 135,000 word forms.
Identifier	406
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://clip.ipipan.waw.pl/Gazetteer
Version	1.0
Last update	2012-07-11

Contacts

--

Agata Savary	
Position	Associate Professor
Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr

Distribution

Availability	Available – restricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

CC-BY-SA		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://clip.ipipan.waw.pl/Gazetteer?action=AttachFile&do=get&target=gazetteer-nkjp-no-pwn.zip http://clip.ipipan.waw.pl/Gazetteer?action=AttachFile&do=get&target=PNEG-LMF-v1.tar.gz	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2012-07-11	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Agata Savary	
	Position	Associate Professor
	Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
	Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-11	

Usage

Foreseen use	NLP applications	
NLP-specific use	Named entity recognition	
Actual uses	NLP applications	
	NLP-specific use	Named entity recognition
	Reports	SAVARY, A., PISKORSKI, J. (2011): Language Resources for Named Entity Annotation in the National Corpus of Polish, to appear in Control and Cybernetics. SAVARY, A., PISKORSKI, J. (2010). Lexicons and Grammars for Named Entity Annotation in the National Corpus of Polish, in Proceedings of the 18th International Conference Intelligent Information Systems (IIS'10), Siedlce, Poland.
	Usage project	National Corpus of Polish

	Project short name	NKJP
	URL	http://www.nkjp.pl
	Funding type	National funds
	Funder	The Polish Ministry of Science and Higher Education (100%)
	Country	Poland
	Start date	2007-12-13
	End date	2011-06-12
Actual use details	Named entity annotation in the National Corpus of Polish. The resource has been used within the Sprout processing platform for pre-annotating named entities.	
NLP applications		
NLP-specific use	Named entity recognition	
Usage project	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
Actual use details	The resource is being used within the NERF tool for automatic named entity recognition in Polish.	

Resource creation

Resource creator	Agata Savary	
	Position	Associate Professor
	Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary

	Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr
	Michal Lenart	
	Position	Programmer
	Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@gmail.com http://zil.ipipan.waw.pl/MichalLenart
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Jakub Piskorski	
	Position	researcher
	Contact	Jana Kazimierza 5 01-248 Warsaw jakub.piskorski@ipipan.waw.pl http://zil.ipipan.waw.pl/JakubPiskorski
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Funding projects	National Corpus of Polish	
	Project short name	NKJP
	URL	http://nkjp.pl/
	Funding type	National funds
	Funder	Polish Ministry of Science and Higher Education
	Country	Poland
	Start date	2007-01-01
	End date	2011-06-30
Creation start date	2004-10-15	

Resource documentation

Reports	SAVARY, A., PISKORSKI, J. (2011): Language Resources for Named Entity Annotation in the National Corpus of Polish, to appear in Control and Cybernetics. More at http://clip.ipipan.waw.pl/Gazetteer .
Tool documentation type	Online

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology Semantics
	Linguistic information	Lemma Case Gender Number Other Part of speech
	Conformance to standards best practices	LMF
Creation	Original source	Publications of the Commission for Standardization of Geographic Names outside Polish Frontiers (Komisja Standaryzacji Nazw Geograficznych poza Granicami Rzeczypospolitej Polskiej), freely available at http://www.gugik.gov.pl/komisja/ . The Polish Wikipedia (http://pl.wikipedia.org) – source of capitals of administrative units of different countries (http://pl.wikipedia.org/wiki/Stolice_jednostek_administracyjnych), rivers (http://pl.wikipedia.org/wiki/Rzeki_Afryki , http://pl.wikipedia.org/wiki/Rzeki_Azji , etc.), historical regions of Europe (http://pl.wikipedia.org/wiki/Kategoria:Regiony_i_krainy_historyczne_Europy), mountain chains (selected from several Wikipedia categories), adjectives and citizen names stemming from country names (http://pl.wiktionary.org/wiki/Indeks:Polski_-_Panstwa_Swiata). The World Gazetteer (http://www.world-gazetteer.com) – source of the list of 200 biggest Polish cities.
	Creation mode	Mixed
	Creation mode details	See section 4.2 in "Lexicons and Grammars for Named Entity Annotation in the National Corpus of Polish".
	Creation tools	SProUT Extraction Platform adapted to Polish Named Entity Annotation with a fully automated rule-based NER system for Polish

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	latin
Size	2.5 mb, 44944 entries, 153477 words	

Character encoding	UTF-8
---------------------------	-------

4.7. LUNA.PL Corpus

General Information

Short name	LUNA.PL
Description	The corpus contains human-human spoken dialogues in Polish. The corpus is annotated on several levels, from transcription of dialogues and their morphosyntactic analysis, to semantic annotation on concepts, predicates and anaphora. Annotation on the morphosyntactic and semantic levels was done automatically and then manually corrected. At the concept level, the annotation scheme comprises about 200 concepts from an ontology designed specially for the project. The set of frames for predicate level annotation was defined as a FrameNet-like resource.
Identifier	407
Resource type	Corpus
URL	http://zil.ipipan.waw.pl/LUNA
Version	1.0
Last update	2011-11-12

Contacts

Malgorzata Marciniak	
Contact	Jana Kazimierza 5 01-248 Warsaw malgorzata.marciniak@ipipan.waw.pl http://zil.ipipan.waw.pl/MalgorzataMarciniak

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

BSD-style	
Restrictions of use	Attribution
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/LUNA?action=AttachFile&do=get&target=LUNA.PL.zip
Fee	free of charge

Metadata

Creation date	2011-10-17

Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Malgorzata Marciniak	
	Contact	Jana Kazimierza 5 01-248 Warsaw malgorzata.marciniak@ipipan.waw.pl http://zil.ipipan.waw.pl/MalgorzataMarciniak
	Michal Lenart	
	Position	Assistant Engineer
	Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@ipipan.waw.pl http://zil.ipipan.waw.pl/MichalLenart
Source	CESAR	

Resource creation

Funding projects	Spoken Language UNderstanding in multilinguAl communication systems	
	Project short name	LUNA
	URL	http://www.ist-luna.eu
	Funding type	EU funds National funds
	Funder	European Commission Polish Ministry of Science and Higher Education
	Country	Poland
	Start date	2006-09-04
	End date	2009-09-03

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL

Size	1.2 gb, 500 files, 12778 utterances, 81049 words	
Annotation	Segmentation	
	Segmentation level	Utterance Word Word group
	Semantic annotation	
	Segmentation level	Clause Word group
Creation	Original source	Warsaw Transport Authority information center recordings
	Creation mode	Mixed
	Creation mode details	Manual transcription of recorded data, automatic creation of files with information about speakers' turns...

4.8. LUNA-WOZ.PL Corpus

General Information

Short name	LUNA-WOZ.PL
Description	The corpus contains human-computer spoken dialogues in Polish. The corpus is annotated on several levels, from transcription of dialogues and their morphosyntactic analysis, to semantic annotation on concepts.
Identifier	408
Resource type	Corpus
URL	http://zil.ipipan.waw.pl/LUNA
Version	1.0
Last update	2011-11-12

Contacts

Malgorzata Marciniak	
Position	Assistant Professor
Contact	Jana Kazimierza 5 01-248 Warsaw
	malgorzata.marciniak@ipipan.waw.pl http://zil.ipipan.waw.pl/MalgorzataMarciniak

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

BSD-style	
Restrictions of use	Attribution
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/LUNA?action=AttachFile&do=get&target=LUNA-WOZ.PL.zip
Fee	free of charge

Metadata

Creation date	2011-10-17	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Michal Lenart	
	Position	Assistant Engineer
	Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@ipipan.waw.pl http://zil.ipipan.waw.pl/MichalLenart
	Malgorzata Marciniak	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw malgorzata.marciniak@ipipan.waw.pl http://zil.ipipan.waw.pl/MalgorzataMarciniak
Source	CESAR	

Resource creation

Funding projects	Spoken Language Understanding in multilinguAl communication systems	
	Project short name	LUNA
	URL	http://www.ist-luna.eu
	Funding type	EU funds National funds
	Funder	European Commission Polish Ministry of Science and Higher Education

	Country	Poland
	Start date	2006-09-04
	End date	2009-09-03

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
Size	140.2 mb, 69 files, 5523 utterances	
Annotation	Segmentation	
	Segmentation level	Utterance Word Word group
	Semantic annotation	
	Segmentation level	Word group
Creation	Original source	Warsaw Transport Authority information center recordings
	Creation mode	Mixed
	Creation mode details	Manual transcription of recorded data, automatic creation of files with information about speakers' turns...

4.9. Morphosyntactic tagset converter for positional tagsets

General Information

Short name	TaCo
Description	TaCo is a statistical morphosyntactic tagset converter designed for positional tagsets, especially Polish tagsets. The typical use is to convert manual annotation of a corpus with tags from one tagset to another tagset. It is based on decision trees produced by C5.0 algorithm and additionally makes use of morphological analyzer Morfeusz. The tool can be configured for converting between various pairs of tagsets and with some additional effort it can be modified to use different morphological analyzers. The converter comes with an example configuration and a trained model for conversion from the IPIPAN Corpus tagset to the National Corpus of Polish tagset.
Identifier	409
Resource type	Tool/service
Tool/service type	Tool
URL	http://zil.ipipan.waw.pl/TaCo
Version	0.1

Last update	2012-06-29
--------------------	------------

Contacts

Bartosz Zaborowski	
Contact	Jana Kazimierza 5 01-248 Warsaw bartosz.zaborowski@ipipan.waw.pl
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – restricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

GPL		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/TaCo?action=AttachFile&do=get&target=taco-0.1.tar.gz	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

--	--

Creation date	2012-06-29	
Metadata creators	Bartosz Zaborowski	
	Contact	Jana Kazimierza 5 01-248 Warsaw bartosz.zaborowski@ipipan.waw.pl
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-06-29	

Usage

Foreseen use	NLP applications
NLP-specific use	Morphosyntactic tagging

Resource creation

Resource creator	Bartosz Zaborowski	
	Contact	Jana Kazimierza 5 01-248 Warsaw bartosz.zaborowski@ipipan.waw.pl
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01

	End date	2013-01-31
Creation start date	2011-02-01	

Resource documentation

Tool documentation type	Manual
--------------------------------	--------

Tool/service

Tool/service type	Tool	
Language dependent	False	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	Linux	
Required software	Ruby (version 1.9.1 or higher) C5.0 (Release 2.07 GPL Edition) Morfeusz (version 0.82)	
Tool/service evaluation	Evaluated	True
	Level	Diagnostic
	Type	Black box
	Criteria	Intrinsic
	Measure	Automatic
	Details	Cross validation of the TaCo tool on the Corpus of Frequency Dictionary of Contemporary Polish, annotated with National Corpus of Polish tagset and IPIPAN Corpus tagset. Conversion achieved 96.1% of correctness (= F-measure = weak correctness).
	Evaluators	Zaborowski Bartosz
		Contact Jana Kazimierza 5 01-248 Warsaw bartosz.zaborowski@ipipan.waw.pl
		Organization Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Tool/service creation	Implementation language	Ruby

	Formalism	C5.0 statistical classifier
--	------------------	-----------------------------

4.10. Spejd

General Information

Short name	Spejd
Description	Spejd is a shallow parser, which allows for simultaneous syntactic parsing and morphological disambiguation.
Identifier	410
Resource type	Tool/service
Tool/service type	Tool
URL	http://zil.ipipan.waw.pl/Spejd
Version	1.3.5
Last update	2012-06-14

Contacts

Bartosz Zaborowski	
Contact	Jana Kazimierza 5 01-248 Warsaw bz233728@students.mimuw.edu.pl
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – unrestricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

GPL	
Restrictions of use	Share alike

Access medium	Downloadable	
Download location	http://sourceforge.net/projects/spejd/files/latest/download	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2012-07-17	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Michal Lenart	
	Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@ipipan.waw.pl http://zil.ipipan.waw.pl/MichalLenart
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-17	

Usage

Foreseen use	NLP applications
---------------------	------------------

NLP-specific use	Parsing	
Actual uses	NLP applications	
	NLP-specific use	Parsing
	Reports	Przepiórkowski A., Bańko M., Górski R. L., Lewandowska-Tomaszczyk B., editors. Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]. PWN Scientific Publishers, Warsaw, 2012.
	Usage project	National Corpus of Polish
		Project short name NKJP
		URL http://www.nkjp.pl
		Funding type National funds
		Funder The Polish Ministry of Science and Higher Education (100%)
		Country Poland
		Start date 2007-12-13
		End date 2011-06-12
	Actual use details	Shallow parsing of the National Corpus of Polish.
	NLP applications	
	NLP-specific use	Parsing
	Reports	Ogrodniczuk M., Przepiórkowski A. Polish Language Processing Chains for Multilingual Information Systems. G. Bouma et al. (ed.): NLDB 2012, LNCS 7337, pp. 152–157. Springer, Heidelberg.
	Usage project	Applied Technology for Language-Aided CMS
		Project short name ATLAS
		URL http://www.atlasproject.eu
		Funding type EU funds National funds
		Funder European Commission (50%) The Polish Ministry of Science and Higher Education (50%)
		Country Poland
		Start date 2010-03-01
		End date 2013-02-28
	Actual use details	Noun phrase detection for the Polish language processing chain.

NLP applications		
NLP-specific use	Parsing	
Reports	Ogrodniczuk M., Kopeć M. Rule-based coreference resolution module for Polish. In Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), pp. 191–200. Faro, Portugal.	
Usage project	Computer-based methods for coreference resolution in Polish texts	
	Project short name	CORE
	URL	http://zil.ipipan.waw.pl/CORE
	Funding type	National funds
	Funder	National Science Centre (100%)
	Country	Poland
	Start date	2011-04-18
	End date	2014-04-17
Actual use details	Noun phrase mention detection for the Polish coreference resolution module.	

Resource creation

Resource creator	Bartosz Zaborowski	
	Contact	Jana Kazimierza 5 01-248 Warsaw bz233728@students.mimuw.edu.pl
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Resource documentation

Reports	Przepiórkowski A. Powierzchniowe przetwarzanie języka polskiego (EN: Shallow parsing of Polish). Warsaw 2008. Akademicka Oficyna Wydawnicza EXIT.
Tool documentation type	Other

Tool/service

Tool/service type	Tool	
Language dependent	False	
Input	Media type	text

	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	Linux Windows	
Required software	Windows XP SP3 or newer (Windows version) GNU/Linux 2.6.9 or newer (Linux version)	
Tool/service creation	Implementation language	C++

4.11. N-grams from balanced National Corpus of Polish

General Information

Short name	N-grams from balanced NKJP
Description	Set of N-grams extracted from balanced National Corpus of Polish for N from 1 to 5. Each unigram is maximum continuous chunk of non-whitespace lower-case characters. The resource contains all unique N-grams followed by number of occurrences.
Identifier	411
Resource type	Corpus
URL	http://zil.ipipan.waw.pl/NKJPNGrams
Version	1.0
Last update	2012-07-01

Contacts

Maciej Ogrodniczuk	
Position	Assistant Professor
Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

BSD-style	
Restrictions of use	Attribution

Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/NKJPNGrams
Fee	free of charge

Metadata

Creation date	2012-07-17	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Michal Lenart	
	Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@ipipan.waw.pl http://zil.ipipan.waw.pl/MichalLenart
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2011-11-26	

Resource creation

Resource creator	Michał Lenart	
	Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@ipipan.waw.pl http://zil.ipipan.waw.pl/MichalLenart
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)

	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31

Corpus text ngram

Media type	textNgram	
Ngram	Base item	Word
	Order	5
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
Size	6.4 gb, 5364398 unigrams, 75395184 bigrams, 170180746 trigrams, 217586930 4 – grams, 232439967 5 – grams	

4.12. Distributable subcorpus of National Corpus of Polish

General Information

Short name	http://zil.ipipan.waw.pl/DistrNKJP
Description	The National Corpus of Polish (PL: Narodowy Korpus Języka Polskiego, NKJP) is a shared initiative of four institutions: Institute of Computer Science at the Polish Academy of Sciences (coordinator), Institute of Polish Language at the Polish Academy of Sciences, Polish Scientific Publishers PWN, and the Department of Computational and Corpus Linguistics at the University of Łódź. It has been registered as a research-development project of the Ministry of Science and Higher Education. The list of sources for the corpus contains classic literature, daily newspapers, specialist periodicals and journals, transcripts of conversations, and a variety of short-lived and internet texts. The resources represent wide diversity with respect to the subject and genre. The spoken part covers both male and female speakers, in various age groups, coming from various regions in Poland. The 1-million subcorpus of NKJP has been manually annotated.
Identifier	412
Resource type	Corpus
URL	http://www.nkjp.pl
Version	1.0

Contacts

Adam Przepiórkowski	
Position	Professor, Head of the Linguistic Engineering Group
Contact	Jana Kazimierza 5 01-248 Warsaw

	adam.przepiorkowski@ipipan.waw.pl http://zil.ipipan.waw.pl/AdamPrzepiorkowski
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – unrestricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

GPL		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/DistrNK.JP	
Fee	free of charge	
Signatories	Adam Przepiórkowski	
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2012-07-05	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Łukasz Degórski	

	Position	Assistant
	Contact	Jana Kazimierza 5 01-248 Warsaw ldegorski@ipipan.waw.pl http://zil.ipipan.waw.pl/LukaszDegorski
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-12-19	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
Size	20.7 gb, 99280766 words	
Annotation	Segmentation	
	Annotation standoff	True
	Segmentation level	Paragraph
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode details	annotated automatically using Pantera tagger
	Annotation tool	Pantera
	Start date	2011-06-06
	End date	2011-06-06
	Segmentation	
	Annotation standoff	True
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI

Annotation mode	Manual
Annotation mode details	annotated automatically using Pantera tagger
Annotation tool	Pantera
Start date	2011-06-06
End date	2011-06-06
Segmentation	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode details	annotated automatically using Pantera tagger
Annotation tool	Pantera
Start date	2011-06-06
End date	2011-06-06
Morphosyntactic annotation - below POS tagging	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode details	annotated automatically using Pantera tagger
Annotation tool	Pantera
Start date	2011-06-06
End date	2011-06-06
Morphosyntactic annotation – POS tagging	
Annotation	True

standoff	
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode details	annotated automatically using Pantera tagger
Annotation tool	Pantera
Start date	2011-06-06
End date	2011-06-06
Syntactic annotation – shallow parsing	
Annotation standoff	True
Segmentation level	Word group
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode details	syntactic words (word-like compounds) and groups output by Spejd
Annotation tool	Spejd
Start date	2011-12-15
End date	2012-01-03
Semantic annotation – named entities	
Annotation standoff	True
Segmentation level	Word group
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Mixed
Annotation mode details	named entities output by Nerf

	Annotation tool	Nerf
	Start date	2011-09-29
	End date	2011-10-04
	Semantic annotation – word senses	
	Annotation standoff	True
	Segmentation level	Word
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Mixed
	Annotation mode details	word sense information output automatically, then manually disambiguated using Anotatornia
	Annotation tool	WSDDE (automatic), Anotatornia (manual)
	Start date	2009-06-29
	End date	2010-12-31
Creation	Original source	the IPI PAN corpus the PELCRA corpus the PWN corpus texts collected by IJP PAN, PELCRA and PWN specifically for NKJP
	Creation tools	Pantera Spejd Nerf

4.13. Morfeusz Polimorf

General Information

Short name	Morfeusz
Description	Morfeusz Polimorf is a variant of the Morfeusz morphological analyser based on Polimorf inflectional dictionary of Polish. Consult the description of Polimorf dictionary for the number of forms and lexemes recognised by the analyser. Morfeusz has the form of a monolithic shared library, which makes it easy to build into client programs with no need for configuration, reading external dictionaries, etc. The dictionary gets compiled to the form of a minimal deterministic finite state automaton, which provides for quick execution and small library size. Morfeusz has been successfully used in several projects both on Windows and on Linux.
Identifier	413
Resource type	Tool/service

Tool/service type	Tool
URL	http://sgjp.pl/morfeusz/index.html
Version	20120730
Last update	2012-07-30

Contacts

Marcin Woliński	
Contact	Jana Kazimierza 5 01-248 Warsaw wolinski@ipipan.waw.pl http://zil.ipipan.waw.pl/MarcinWolinski
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – restricted use	
IPR holder	Marcin Woliński	
	Contact	Jana Kazimierza 5 01-248 Warsaw wolinski@ipipan.waw.pl http://zil.ipipan.waw.pl/MarcinWolinski
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Licences

BSD-style		
Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://sgjp.pl/morfeusz/dopobrania.html	
Fee	free of charge	
Distribution rights holder	Marcin Woliński	
	Contact	Jana Kazimierza 5 01-248 Warsaw wolinski@ipipan.waw.pl http://zil.ipipan.waw.pl/MarcinWolinski
	Organization	Institute of Computer Science, Polish Academy of Sciences

		Linguistic Engineering Group ipi@ipipan.waw.pl
--	--	--

Metadata

Creation date	2012-07-25	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Marcin Woliński	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw wolinski@ipipan.waw.pl http://zil.ipipan.waw.pl/MarcinWolinski
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-25	

Usage

Foreseen use	NLP applications	
NLP-specific use	Morphological analysis	
Actual uses	NLP applications	
	NLP-specific use	Morphological analysis
	Reports	Used for automated morphological analysis of the IPI PAN Corpus.
	Usage project	IPI PAN Corpus
		Project short name IPI PAN Corpus

	URL	http://korpus.pl/index.php?lang=en&page=welcome	
	Funding type	National funds	
	Funder	State Committee for Scientific Research (100%)	
	Country	Poland	
	Start date	2001-04-01	
	End date	2004-03-31	
Actual use details	Morphological analysis of the entire IPI PAN Corpus.		
NLP applications			
NLP-specific use	Morphological analysis		
Reports	Used for automated morphological analysis of the National Corpus of Polish.		
Usage project	National Corpus of Polish		
	Project short name	NKJP	
	URL	http://www.nkjp.pl	
	Funding type	National funds	
	Funder	The Polish Ministry of Science and Higher Education (100%)	
	Country	Poland	
	Start date	2007-12-13	
	End date	2011-06-12	
Actual use details	Morphological analysis of the entire National Corpus of Polish.		
NLP applications			
NLP-specific use	Morphological analysis		
Reports	Ogrodniczuk M., Przepiórkowski A. Polish Language Processing Chains for Multilingual Information Systems. G. Bouma et al. (ed.): NLDB 2012, LNCS 7337, pp. 152–157. Springer, Heidelberg.		
Usage project	Applied Technology for Language-Aided CMS		
	Project short name	ATLAS	
	URL	http://www.atlasproject.eu	
	Funding type	EU funds National funds	

	Funder	European Commission (50%) The Polish Ministry of Science and Higher Education (50%)
	Country	Poland
	Start date	2010-03-01
	End date	2013-02-28
Actual use details	Morphological analysis of the National Corpus of Polish. Tool trained on the manually annotated million-word subcorpus has been used to annotate the entire NKJP corpus.	
NLP applications		
NLP-specific use	Morphological analysis	
Reports	Ogrodniczuk M., Kopeć M. Rule-based coreference resolution module for Polish. In Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), pp. 191–200. Faro, Portugal.	
Usage project	Computer-based methods for coreference resolution in Polish texts	
	Project short name	CORE
	URL	http://zil.ipipan.waw.pl/CORE
	Funding type	National funds
	Funder	National Science Centre (100%)
	Country	Poland
	Start date	2011-04-18
	End date	2014-04-17
Actual use details	Morphological analysis for the Polish coreference resolution module.	

Resource creation

Resource creator	Marcin Woliński	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw wolinski@ipipan.waw.pl http://zil.ipipan.waw.pl/MarcinWolinski
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Creation start date	2001-09-01	

Resource documentation

Reports	Woliński M. 2006. Morfeusz – a practical tool for the morphological analysis of Polish. In: Mieczysław A. Kłopotek, Sławomir T. Wierchoń and Krzysztof Trojanowski, editors, Proceedings of the International Intelligent Information Systems: Intelligent Information Processing and Web Mining'06 Conference, pages 511–520, Wisła, Poland, June. See also http://sgjp.pl/morfeusz/morfeusz.html .
Tool documentation type	Other

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Text/plain	
	Media type	text
	Modality type	Written language
Output	Lemmatization	
	Media type	text
	Modality type	Written language
	Tagset	Morphological tagset in the IPI PAN corpus, see http://www.ipipan.waw.pl/~wolinski/publ/znakowanie.pdf (published in Polonica XXII/XXIII, 2003, pp. 39-55 in Polish).
	Segmentation level	Word
Operating system	Linux Mac OS Windows	
Tool/service creation	Implementation language	C++
	Formalism	FSA

4.14. Morfologik Inflectional Dictionary

General Information

Short name	Morfologik
Description	The morphological dictionary for Polish, a special release of PoliMorf dictionary in the format compatible with the previous dictionary Morfologik. Morfologik is an open-source morphological dictionary of Polish. It contains 216,992 lexemes and 3,475,809 word forms. The dictionary was created by enriching the Polish ispell/hunspell dictionary with morphological information, which was possible thanks to the structure of the original dictionary that retained important grammatical distinctions. The process of conversion relied

	on a series of scripts, and the resulting dictionary was later augmented with manually entered information. Unfortunately, the original source dictionary did not contain sufficient structure to allow reliable detection of some information, such as the exact subgender of the masculine for substantives. This information was added manually and using heuristic methods, however its reliability is low. Considering the fact that the substantives are about one third of the dictionary content (and almost half of them are masculine), this limitation is severe. The tagset of the dictionary is inspired by the IPI PAN Tagset. However, Morfologik diverges from that tagset and from Morfeusz, as it never splits orthographic ("space-to-space") words into smaller dictionary words (i.e. so-called agglutination is not considered). Moreover, due to the lack of information in the ispell dictionary, some forms are not completely annotated, and are marked as irregular. There is, however, some additional mark up added to reflexive verbs, which is not present in the original IPI PAN Tagset. This was introduced for the purposes of the grammar checker LanguageTool that used the dictionary extensively.
Identifier	414
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://sourceforge.net/projects/morfologik
Version	2.0

Contacts

Marcin Milkowski	
Position	Assistant Professor
Contact	Nowy Świat 72 00-330 Warsaw marcin.milkowski@ifispan.waw.pl http://zil.ipipan.waw.pl/MarcinMilkowski
Organization	Institute of Philosophy and Sociology, Polish Academy of Sciences Department of Logic and Cognitive Science secretar@ifispan.waw.pl

Distribution

Availability	Available – unrestricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

BSD-style		
Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://sourceforge.net/projects/morfologik/files/morfologik/	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2012-07-24	
Metadata creators	Marcin Milkowski	
	Position	Assistant Professor
	Contact	Nowy Świat 72 00-330 Warsaw marcin.milkowski@ifispan.waw.pl http://zil.ipipan.waw.pl/MarcinMilkowski
	Organization	Institute of Philosophy and Sociology, Polish Academy of Sciences Department of Logic and Cognitive Science secretar@ifispan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-24	

Resource creation

Resource creator	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5

		01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2011-02-01	

Resource documentation

Reports	Marcin Milkowski. Developing an open-source, rule-based proofreading tool. In: Software: Practice & Experience 40 (7), pp. 543-566. http://doi.wiley.com/10.1002/spe.971
----------------	---

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology
	Linguistic information	Lemma Aspect Auxiliary Case Degree Gender Inflection Mood Number Person Tense Part of speech
Creation	Original source	Morfeusz SGJP Morfologik

	Creation mode	Mixed
	Creation tools	Kuźnia

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	latin
Size	6382227 entries	
Character encoding	UTF-8	

4.15. Grammatical Lexicon of Polish Phraseology

General Information

Short name	SEJF
Description	The Grammatical Lexicon of Polish Phraseology (Słownik elektroniczny języka polskiego dla wyrażen frazeologicznych) is an electronic lexicon containing multi-word units (mainly nominal, adjectival and adverbial compounds) of the general (non-terminological) Polish language. It has been created within the ERDF Nekt project and contains about 5,000 multi-word lexemes, 93,000 corresponding inflected forms, and 160 graph-based inflection paradigms.
Identifier	415
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://zil.ipipan.waw.pl/SEJF
Version	1.0
Last update	2012-07-23

Contacts

Monika Czerepowicka	
Position	Associate Professor
Contact	ul. Kurta Obizta 1 10-725 Olsztyn czerepowicka@gmail.com http://www.uwm.edu.pl/polonistyka/index.php?

	option=com_content&view=article&id=95&catid=50&Itemid=9
Organization	University of Warmia and Mazury in Olsztyn Instytut Filologii Polskiej filpol.human@uwm.edu.pl

Distribution

Availability	Available – restricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

CC-BY-SA		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/SEJF?action=AttachFile&do=get&target=SEJF.tar.gz	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2012-07-11	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl

	http://zil.ipipan.waw.pl/MaciejOgrodniczuk
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Agata Savary	
Position	Associate Professor
Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr
Source	CESAR
Metadata language ID	en
Metadata last date updated	2012-07-18

Usage

Foreseen use	NLP applications
NLP-specific use	Morphological analysis Morphosyntactic tagging Parsing

Resource creation

Resource creator	Agata Savary	
	Position	Associate Professor
	Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
	Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr
	Monika Czerepowicka	
	Position	Associate Professor
	Contact	ul. Kurta Obizta 1 10-725 Olsztyn czerepowicka@gmail.com

		http://www.uwm.edu.pl/polonistyka/index.php?option=com_content&view=article&id=95&catid=50&Itemid=9
	Organization	University of Warmia and Mazury in Olsztyn Instytut Filologii Polskiej filpol.human@uwm.edu.pl
Funding projects	NEKST — Adaptive system supporting solving problems on the basis of content analysis of electronic documents	
	Project ID	POIG.01.01.02-14-013/09
	Funding type	EU funds National funds
	Country	Poland
	Start date	2009-04-01
	End date	2014-02-10
Creation start date	2010-03-19	

Resource documentation

Reports	GRALIŃSKI, F., SAVARY, A., CZEREPOWICKA, M., MAKOWIECKI, F. (2010): Computational Lexicography of Multi-Word Units: How Efficient Can It Be?, in Proceedings of Multiword Expressions: from Theory to Applications (MWE 2010), Workshop at COLING 2010, Beijing, China, August 28. More at http://zil.ipipan.waw.pl/SEJF .
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology Syntax
	Linguistic information	Lemma Aspect Auxiliary Case Degree Gender Inflection Mood Number Person Tense Other Part of speech
Creation	Original source	KOSEK, I. (2008): "Fleksja i składnia nieciągłych imiennych jednostek leksykalnych", Olsztyn. BAŃKO, M. (2004): "Słownik porównań",

	Wydawnictwo Naukowe PWN, Warszawa. CZEREPOWICKA, M. (2006): "Opis powierzchniowskładniowy wyrażen niestandardowych typu na lewo, do dziś, po trochu, na zawsze we współczesnym języku polskim", Akademicka Oficyna Wydawnicza EXIT, Warszawa. WOJDAK, P. (2004): "Przysłowki polisegmentalne w modelu składniowym polszczyzny", Wydawnictwo Naukowe US, Szczecin. pIWordNet (http://plwordnet.pwr.wroc.pl/wordnet)
Creation mode	Mixed
Creation mode details	see "Computational Lexicography of Multi-Word Units: How Efficient Can It Be?"
Creation tools	Toposław (http://zil.ipipan.waw.pl/Toposlaw) Morfeusz (http://sgjp.pl/morfeusz/) Multiflex (http://www.springerlink.com/content/n265j22n73084433/) graph editor from Unitex (http://igm.univ-mlv.fr/~unitex/)

Texts

Media type	text
Linguality type	Monolingual
Languages	Polish
	Language ID PL
	Language script latin
Size	0.4 mb, 3176 entries, 67879 multi-word units, 159 rules
Character encoding	UTF-8

4.16. Grammatical Lexicon of Polish Economical Phraseology

General Information

Short name	SEJFEK
Description	The Grammatical Lexicon of Polish Economical Phraseology (SEJFEK – Słownik Elektroniczny Języka polskiego dla wyrażen frazeologicznych z EKonomii) is an electronic lexicon containing multi-word nominal terms of Polish economical and financial terminology. It has been created within the ERDF Nekst project.
Identifier	416
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://zil.ipipan.waw.pl/SAWA
Version	1.0
Last update	2012-07-19

Contacts

Agata Savary	
Position	Associate Professor
Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr

Distribution

Availability	Available – restricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

CC-BY-SA		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/SAWA?action=AttachFile&do=get&target=SEJFEK.tar.gz	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2012-07-19	
Metadata creators	Agata Savary	
	Position	Associate Professor
	Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
	Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-20	

Usage

Foreseen use	NLP applications
NLP-specific use	Morphological analysis Morphosyntactic tagging Parsing

Resource creation

Resource creator	Agata Savary	
	Position	Associate Professor
	Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
	Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr
	Filip Makowiecki	
	Contact	f.makowiecki@gmail.com
	Organization	University of Warsaw f.makowiecki@gmail.com
Funding projects	NEKST — Adaptive system supporting solving problems on the basis of content analysis of electronic documents	
	Project ID	POIG.01.01.02-14-013/09

	Funding type	EU funds National funds
	Country	Poland
	Start date	2009-04-01
	End date	2014-02-10
Creation start date	2009-04-01	

Resource documentation

Reports	GRALIŃSKI, F., SAVARY, A., CZEREPOWICKA, M., MAKOWIECKI, F. (2010): Computational Lexicography of Multi-Word Units: How Efficient Can It Be?, in Proceedings of Multiword Expressions: from Theory to Applications (MWE 2010), Workshop at COLING 2010, Beijing, China, August 28.
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology Syntax
	Linguistic information	Lemma Aspect Auxiliary Case Degree Gender Inflection Mood Number Person Tense Other Part of speech
Creation	Original source	GRALIŃSKI, F., SAVARY, A., CZEREPOWICKA, M., MAKOWIECKI, F. (2010): Computational Lexicography of Multi-Word Units: How Efficient Can It Be?, in Proceedings of Multiword Expressions: from Theory to Applications (MWE 2010), Workshop at COLING 2010, Beijing, China, August 28.
	Creation mode	Mixed
	Creation tools	Toposław (http://zil.ipipan.waw.pl/Toposlaw/) Morfeusz (http://sgjp.pl/morfeusz/) Multiflex (http://www.springerlink.com/content/n265j22n73084433/) graph editor from Unitex (http://igm.univ-mlv.fr/~unitex/)

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	latin
Size	0.014 mb, 11212 entries, 146861 multi-word units, 305 rules	
Character encoding	UTF-8	

4.17. Grammatical Lexicon of Warsaw Urban Proper Names

General Information

Short name	SAWA
Description	The Grammatical Lexicon of Warsaw Urban Proper Names (SAWA - Słownik elektroniczny nazewnictwa Warszawy) is an electronic lexicon containing about 9,000 proper names of places related to the Warsaw transportation system, i.e. names of streets, squares, monuments, buildings, bus, tram and subway stops, etc., as well as names of persons to whom some objects (notably streets) are dedicated. Previous names (notably those used before 1989) are also included. Their morphosyntax is described by over 450 graph-based inflection paradigms, which allow an automatic generation of over 300,000 inflectional and syntactic variants. It has been developed within a French-Polish Polonium project and within nationally funded Polish project.
Identifier	417
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://zil.ipipan.waw.pl/SAWA
Version	1.0
Last update	2012-07-19

Contacts

Malgorzata Marciniak	
Position	Associate Professor
Contact	Jana Kazimierza 5 01-248 Warsaw mm@ipipan.waw.pl http://zil.ipipan.waw.pl/MalgorzataMarciniak
Organization	Institute of Computer Science, Polish Academy of Sciences

	Linguistic Engineering Group ipi@ipipan.waw.pl
--	--

Distribution

Availability	Available – restricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

CC-BY-SA		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/SAWA?action=AttachFile&do=get&target=SAWA.tar.gz	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2012-07-19	
Metadata creators	Agata Savary	
	Position	Associate Professor
	Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
	Organization	Université François Rabelais Tours

	Laboratoire d'Informatique li@univ-tours.fr
Source	CESAR
Metadata language ID	en
Metadata last date updated	2012-07-20

Usage

Foreseen use	NLP applications
NLP-specific use	Morphological analysis Morphosyntactic tagging Parsing

Resource creation

Resource creator	Agata Savary	
	Position	Associate Professor
	Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
	Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr
	Malgorzata Marciniak	
	Position	Associate Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw mm@ipipan.waw.pl http://zil.ipipan.waw.pl/MalgorzataMarciniak
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Celina Heliasz	
	Contact	celina.heliasz@uw.edu.pl
	Organization	University of Warsaw Formal Linguistics Department jsbien@uw.edu.pl a.s.boguslawski@uw.edu.pl m.m.danielewicz@uw.edu.pl j.z.wajszczuk@uw.edu.pl

	Joanna Rabiega-Wiśniewska	
	Contact	jwr@cereza.pl
	Piotr Sikora	
	Contact	piotr.sikora@student.uw.edu.pl
	Marcin Woliński	
	Position	Associate Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw wolinski@ipipan.waw.pl http://www.ipipan.waw.pl/~wolinski/
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Funding projects	Description morphologique de noms propres polonais pour applications multilingues	
	Funding type	National funds
	Country	Poland France
	Start date	2007-01-01
	End date	2008-12-31
	Spoken language understanding in multilingual communication systems	
	Funding type	National funds
	Country	Poland
	Start date	2008-01-01
	End date	2009-12-31
Creation start date	2007-01-01	

Resource documentation

Reports	SAVARY, A., RABIEGA-WIŚNIEWSKA, J., WOLIŃSKI, M. (2009): "Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex", in MARCINIAK, M., MYKOWIECKA, A. (eds.) "Aspects of Natural Language Processing", Lecture Notes in Computer Science 5070, Springer Verlag, pp. 111-141. More at http://zil.ipipan.waw.pl/SAWA .
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology Syntax

	Linguistic information	Lemma Aspect Auxiliary Case Degree Gender Inflection Mood Number Person Tense Other Part of speech
Creation	Original source	Savary, A., Rabiega-Wisniewska, J., Woliński, M. (2009): "Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex", in Marciniak, M., Mykowiecka, A. (eds.) "Aspects of Natural Language Processing", Lecture Notes in Computer Science 5070, Springer Verlag, pp. 111-141. Marciniak, M., Rabiega-Wisniewska, J., Savary, A., Woliński, M., Heliasz, C. (2009): "Constructing an Electronic Dictionary of Polish Urban Proper Names", in Recent Advances in Intelligent Information Systems (Proceedings of the Balto-Slavonic Natural Language Processing Workshop, Kraków), Academic Publishing House EXIT, Warsaw, pp. 743-749.
	Creation mode	Mixed
	Creation tools	Toposław (http://zil.ipipan.waw.pl/Toposlaw/) Morfeusz (http://sgjp.pl/morfeusz/) Multiflex (http://www.springerlink.com/content/n265j22n73084433/) graph editor from Unitex (http://igm.univ-mlv.fr/~unitex/)

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	latin
Size	1.5 mb, 9000 entries, 300000 words, 450 rules	
Character encoding	UTF-8	

4.18. Multilingual lexicon of toponyms

General Information

Short name	WikiTopoPl
Description	The multilingual lexicon of toponyms (WikiTopoPl) contains a list of over 155,000 polish geographical proper names (countries, cities, regions, hydronyms, etc) and their equivalents in Bulgarian, German, modern Greek, English, Croatian, Hungarian, Romanian, Slovak and Serbian. These data (whenever available) have been automatically extracted from the open encyclopedia Wikipedia. The Wikipedia categories attached to the lexicon entries have been mapped to a short list of succinct categories compliant with Prolexbase, a multilingual ontology of proper names.
Identifier	418
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Lexicon
URL	http://bach.ipipan.waw.pl/redmine/issues/227
Version	0.2
Last update	2012-10-25

Contacts

Leszek Manicki	
Contact	Wilczak 13/54 61-623 Poznań lebiega@gmail.com
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – restricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

CC-BY-SA	
Restrictions of use	Share alike

Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/WikiTopoPl?action=AttachFile&do=view&target=WikiTopoPl.tar.gz	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2012-07-24	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Leszek Manicki	
	Contact	Wilczak 13/54 61-623 Poznań lebiega@gmail.com
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-10-30	

Usage

Foreseen use	NLP applications

NLP-specific use	Machine translation
	Named entity recognition

Resource creation

Resource creator	Leszek Manicki	
	Contact	Wilczak 13/54 61-623 Poznań lebiega@gmail.com
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
	Creation start date	2012-05-15

Lexical conceptual resource

Lexical conceptual resource type	Lexicon	
Lexical conceptual resource encoding	Encoding level	Other Semantics
	Linguistic information	Lemma Semantics – semantic class Translation equivalent

Texts

Media type	text
Linguality type	Multilingual

Languages

Polish	
Language ID	PL
Language script	Latin
Size	155000 entries
Bulgarian	
Language ID	BG
Language script	Cyrillic
Size	8000 entries
German	
Language ID	DE
Language script	Latin
Size	43000 entries
Modern Greek	
Language ID	EL
Language script	Greek
Size	3000 entries
English	
Language ID	EN
Language script	Latin
Size	155000 entries
Croatian	
Language ID	HR
Language script	Latin
Size	4375 entries
Hungarian	
Language ID	HU
Language script	Latin
Size	16000 entries
Romanian	
Language ID	RO
Language	Latin

	script	
	Size	19000 entries
	Slovak	
	Language ID	SK
	Language script	Latin
	Size	12000 entries
	Serbian	
	Language ID	SR
	Language script	Cyrillic
	Size	21000 entries
Modality	Modality type	Written language
Size	1.5 mb, 155000 entries	
Character encoding	UTF-8	

4.19. Polish Valence Dictionary

General Information

Short name	Walenty
Description	The Polish Valence Dictionary (Walenty) contains a description of argument structures of 1774 Polish verbs and quasi-verbal predicates. The entries are represented through a number of individual frames, each frame corresponding to a set of positions which may be filled by phrases of appropriate types and parameters. Individual positions may be marked for their status as a subject or a passivisable object, and for their role in control relations with other positions in the argument structure. The Polish Valence Dictionary is an adaptation of the Syntactic Dictionary of Polish Verbs (Świdziński 1994) in a digitised version expanded by Witold Kieraś to include a number of frequent verbs missing from the original dictionary. In addition to expanding the number of frames, the presented resource includes information about new features, including sentential subjects, passivisation, control relations, semantic categories of adverbial phrases, and possibility of coordination of different types of arguments.
Identifier	419
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://clip.ipipan.waw.pl/Walenty

Contacts

Filip Skwarski	

Contact	Jana Kazimierza 5 01-248 Warsaw filip.skwarski@ipipan.waw.pl http://zil.ipipan.waw.pl/FilipSkwarski
----------------	---

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

CC-BY-SA	
Restrictions of use	Attribution Share alike
Access medium	Downloadable
Download location	http://clip.ipipan.waw.pl/Walenty?action=AttachFile&do=view&target=polish_valence_dictionary_v2.zip
Fee	free of charge

Metadata

Creation date	2012-07-20	
Metadata creators	Filip Skwarski	
	Contact	Jana Kazimierza 5 01-248 Warsaw filip.skwarski@ipipan.waw.pl http://zil.ipipan.waw.pl/FilipSkwarski
Source	CESAR	

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Parsing	
Actual uses	NLP applications	
	Actual use details	Recognition of verbal complements by parsers.

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Creation	Original source	Świdziński M. Syntactic Dictionary of Polish Verbs. Uniwersytet

		Warszawski / Universiteit van Amsterdam, 1994.
	Creation mode	Mixed
	Creation tools	Slowal - web application designed for manual edition of valence frames (see http://zil.ipipan.waw.pl/Slowal)

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
Size	0.03 mb, 1774 entries	

4.20. Summarizer

General Information

Short name	Summarizer
Description	Summarizer is a tool for creating short text summaries. It utilises text extraction method, i.e. the output consists of sentences from the original text. The tool uses a number of machine learning algorithms, including neural networks, linear regression, Bayesian networks and decision trees. The output sentences are chosen based on different signals, such as the length of the sentence, its position in the text structure and properties of the words it contains. The system was trained specifically for newspaper articles in Polish. It is possible, however, to adjust it for other kinds of documents and languages.
Identifier	420
Resource type	Tool/service
Tool/service type	Tool
URL	http://clip.ipipan.waw.pl/Summarizer

Contacts

Joanna Świetlicka	
Contact	44 Reynolds House, Erasmus Street SW1P 4HP London j.swietlicka@gmail.com

Distribution

Availability	Available – restricted use	
IPR holder	Joanna Świetlicka	
	Contact	44 Reynolds House, Erasmus Street

		SW1P 4HP London j.swietlicka@gmail.com
--	--	---

Licences

CC-BY	
Restrictions of use	Attribution
Access medium	Downloadable
Download location	http://clip.ipipan.waw.pl/Summarizer
Fee	free of charge

Metadata

Creation date	2012-07-06	
Metadata creators	Joanna Świetlicka	
	Contact	44 Reynolds House, Erasmus Street SW1P 4HP London j.swietlicka@gmail.com
Metadata language ID	en	
Metadata last date updated	2012-07-06	

Resource creation

Resource creator	Joanna Świetlicka	
	Contact	44 Reynolds House, Erasmus Street SW1P 4HP London j.swietlicka@gmail.com
Creation start date	2010-01-01	

Resource documentation

Reports	Świetlicka J. Machine learning methods in automated text summarization. MSc thesis, 2010.
----------------	---

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Media type	text

	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Tool/service creation	Implementation language	Java

4.21. morfologik-stemming

General Information

Short name	morfologik-stemming
Description	Morfologik-stemming is a library featuring morphological analysis, spelling correction, and building of finite-state automata for these purposes. It is bundled with a morphological dictionary for Polish, Morfologik.
Identifier	421
Resource type	Tool/service
Tool/service type	Tool
URL	http://sourceforge.net/projects/morfologik
Version	20120730
Last update	2012-07-30

Contacts

Dawid Weiss	
Contact	Bożnicza 11/57 61-751 Poznań info@carrotsearch.com http://www.carrotsearch.com
Organization	Carrot Search info@carrotsearch.com

Distribution

Availability	Available – restricted use	
IPR holder	Dawid Weiss	
	Contact	Bożnicza 11/57 61-751 Poznań info@carrotsearch.com http://carrotsearch.com/about.html
	Organization	Carrot Search info@carrotsearch.com

Licences

BSD-style		
Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://sourceforge.net/projects/morfologik/files/morfologik-stemming/	
Fee	free of charge	
Distribution rights holder	Dawid Weiss	
	Contact	Bożnicza 11/57 61-751 Poznań info@carrotsearch.com http://carrotsearch.com/about.html
	Organization	Carrot Search info@carrotsearch.com

Metadata

Creation date	2012-07-26	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Marcin Milkowski	
	Position	Assistant Professor
	Contact	Nowy Świat 72 00-330 Warsaw mmilkows@ifspan.waw.pl http://zil.ipipan.waw.pl/MarcinMilkowski
	Organization	Institute of Philosophy and Sociology, Polish Academy of Sciences Department of Logic and Cognitive Science secretar@ifspan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date	2012-07-25	

updated	
---------	--

Usage

Foreseen use	NLP applications	
NLP-specific use	Morphological analysis Spell checking	
Actual uses	NLP applications	
	NLP-specific use	Morphological analysis
	Reports	Used for automated morphological analysis in LanguageTool.
	Usage project	LanguageTool
		Project short name LanguageTool
		URL http://www.languagetool.org
		Funding type Other
	Actual use details	Morphological analysis in LanguageTool.
	NLP applications	
	NLP-specific use	Morphological analysis
	Reports	Used for automated morphological analysis in LanguageTool.
	Usage project	LanguageTool
		Project short name LanguageTool
		URL http://www.languagetool.org
		Funding type Other
	Actual use details	Morphological analysis in LanguageTool.

Resource creation

Resource creator	Dawid Weiss	
	Position	Owner
	Contact	Bożnicza 11/57 61-751 Poznań info@carrotsearch.com http://carrotsearch.com/about.html
	Organization	Carrot Search

	info@carrotsearch.com
Creation start date	2006-08-17

Resource documentation

Reports	Milkowski, Marcin. 2010. "Developing an Open-source, Rule-based Proofreading Tool." Software: Practice and Experience 40 (7): 543–566. doi:10.1002/spe.971. http://doi.wiley.com/10.1002/spe.971 . See also http://morfologik.blogspot.com .
Tool documentation type	Other

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Text/plain	
	Media type	text
	Modality type	Written language
Output	Lemmatization	
	Media type	text
	Modality type	Written language
	Tagset	A special flatten representation of the morphological tagset in the IPI PAN corpus, without intra-word segmentation.
	Segmentation level	Word
Operating system	Linux Mac OS Windows	
Tool/service creation	Implementation language	Java
	Formalism	FSA

4.22. Corpus of the Polish language of the 1960s

General Information

Short name	PL196x
Description	The Corpus of the Polish language of the 1960s (originally: the corpus of frequency dictionary of contemporary Polish) was prepared to create a general frequency dictionary of contemporary Polish. The work started in 1967 with partial results published in 1972-1977 and the completed dictionary in 1990. The corpus was later augmented in various respects, both by manual editing and automated procedures. Corpus data contain 10,000

	samples divided into 5 parts: essays, news, scientific texts, fiction and plays. Every sample is approximately 50 words long, they all come from texts published between 1963 and 1967 and contain bibliographic description of its source. Each word is tagged with its base form and some morphological properties. Sentence boundaries are also marked.
Identifier	422
Resource type	Corpus
URL	http://clip.ipipan.waw.pl/PL196x
Version	1.0

Contacts

Maciej Ogrodniczuk	
Position	Assistant Professor
Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – unrestricted use	
IPR holder	Ida Kurecz	
	Contact	un@known.pl
	Andrzej Lewicki	
	Contact	un@known.pl
	Jadwiga Sambor	
	Contact	un@known.pl
	Krzysztof Szafran	
	Contact	k.szafran@mimuw.edu.pl
	Pawel Woronczak	
	Contact	jpawelw@uni.wroc.pl
	Lucyna Woronczakowa	
	Contact	jpawelw@uni.wroc.pl

Licences

GPL	
Restrictions of use	Attribution

Access medium	Downloadable	
Download location	http://clip.ipipan.waw.pl/PL196x	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2012-07-06	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-06	

Usage

Foreseen use	Human use
NLP-specific use	Linguistic research
Foreseen use	NLP applications
NLP-specific use	Linguistic research

Resource documentation

Reports	Kurcz, Ida; Lewicki, Andrzej; Sambor, Jadwiga; Szafran, Krzysztof; Woronczak, Jerzy. Polish language in the sixties (in English, introduction to the printed edition of the frequency dictionary). See http://clip.ipipan.waw.pl/PL196x for more publications in Polish.
----------------	---

Texts

--	--

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
Size	80.2 mb, 10000 texts, 500000 words	
Annotation	Segmentation	
	Annotation standoff	True
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Manual
	Annotation mode details	annotated manually at frequency dictionary preparation, further verified in 2012 using Anotatornia tool
	Annotation tool	Anotatornia
	Start date	2009-06-29
	End date	2010-09-30
	Segmentation	
	Annotation standoff	True
	Segmentation level	Word
	Format	text/xml
	Tagset	NKJP tagset
	Conformance to standards best practices	TEI
	Annotation mode	Manual
	Annotation mode details	tokens (word-like segments – see documentation of Morfeusz SGJP for details); segmentation revised by annotators using Anotatornia
	Annotation tool	Anotatornia
	Start date	2009-06-29
	End date	2010-09-30
	Lemmatization	
	Annotation	True

standoff	
Segmentation level	Word
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Manual
Annotation mode details	lemma variants available in the original corpus, in 2012 manually verified using Anotatornia
Annotation tool	Anotatornia
Start date	2009-06-29
End date	2010-09-30
Morphosyntactic annotation - below POS tagging	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode	Manual
Annotation mode details	MSD tags conformant to the tagset of the original corpus have been converted in 2012 to NKJP tagset using Anotatornia
Annotation tool	Anotatornia
Start date	2009-06-29
End date	2010-09-30
Morphosyntactic annotation – POS tagging	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI

	Annotation mode	Manual
	Annotation mode details	POS tags (CTAGs) conformant to the tagset of the original corpus have been converted in 2012 to NKJP tagset using Anotatoria
	Annotation tool	Anotatoria
	Start date	2009-06-29
	End date	2010-09-30
Creation	Original source	essays, news, scientific texts, fiction and plays
	Creation mode	Manual
	Creation mode details	Texts have been collected and sampled by dictionary authors.

4.23. Shallow Grammar for the National Corpus of Polish

General Information

Short name	NKJPGrammar
Description	Shallow Grammar for the National Corpus of Polish is a set of rules which was used for the automatic pre-annotation of the National Corpus of Polish at the syntactic level. It was constructed manually and encoded in the shallow parsing system Spejd (http://nlp.ipipan.waw.pl/Spejd/). It consists of 1187 rules for multiword entities, abbreviations, syntactic words, and syntactic groups.
Identifier	423
Resource type	Language description
Language description type	Grammar
URL	http://clip.ipipan.waw.pl/LRT?action=AttachFile&do=view&target=gramatyka_Spejd_NKJP_1.0.zip
Version	1.0
Last update	2012-07-24

Contacts

Katarzyna Głowińska	
Position	grammar author
Contact	Jana Kazimierza 5 01-248 Warsaw k.glowinska@gmail.com

Distribution

--	--

Availability	Available – restricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

GPL		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://clip.ipipan.waw.pl/LRT?action=AttachFile&do=view&target=gramatyka_Spejd_NKJP_1.0.zip	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2012-07-24	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Katarzyna Głowińska	
	Contact	Jana Kazimierza 5 01-248 Warsaw k.glowinska@gmail.com

Source	CESAR
Metadata language ID	en
Metadata last date updated	2012-07-24

Usage

Foreseen use	NLP applications	
NLP-specific use	Morphosyntactic tagging	
Actual uses	NLP applications	
	NLP-specific use	Named entity recognition
	Reports	Głowińska K., Przepiórkowski A. The Design of Syntactic Annotation Levels in the National Corpus of Polish. In: LREC 2010 proceedings. Waszczuk, J., Głowińska, K., Savary, A., Przepiórkowski, A. Tools and Methodologies for Annotating Syntax and Named Entities in the National Corpus of Polish. In: Proceedings of Computational Linguistics – Applications (CLA 2010), Workshop at IMCSIT 2010, Wisła, Poland, October 18-20.
	Usage project	National Corpus of Polish
		Project short name NKJP
		URL http://www.nkjp.pl
		Funding type National funds
		Funder The Polish Ministry of Science and Higher Education (100%)
		Country Poland
		Start date 2007-12-13
		End date 2011-06-12
	Actual use details	Syntactic annotation in the National Corpus of Polish. Spejd grammar was used for the automatic pre-annotation of one-million-word subcorpus. Then, after some improvements, was used to annotate the entire NKJP corpus.

Resource creation

Resource creator	Katarzyna Głowińska	
	Contact	Jana Kazimierza 5 01-248 Warsaw k.glowinska@gmail.com
Funding projects	National Corpus of Polish	

	Project short name	NKJP
	URL	http://nkjp.pl/
	Funding type	National funds
	Funder	Polish Ministry of Science and Higher Education
	Country	Poland
	Start date	2007-01-01
	End date	2011-06-30
Creation start date	2010-02-01	

Resource documentation

Reports	<p>Waszczuk, J., Głowińska, K., Savary, A., Przepiórkowski, A. Tools and Methodologies for Annotating Syntax and Named Entities in the National Corpus of Polish. In: Proceedings of Computational Linguistics – Applications (CLA 2010), Workshop at IMCSIT 2010, Wisła, Poland, October 18-20.</p> <p>Głowińska K., Przepiórkowski A. The Design of Syntactic Annotation Levels in the National Corpus of Polish. W: LREC 2010 proceedings.</p>
----------------	---

Language description

Language description type	Grammar		
Language description encoding	Encoding level	Syntax	
	Theoretic model	Shallow approach to syntactic analysis. Spejd grammar consists of rules for combining words into constituents at the level of syntactic words and syntactic groups. At the former, fine-grained word-level tokens are replaced by coarse-grained syntactic words, i.e., traditional word forms, including analytical tense and mood forms, reflexive verbs, discontinuous conjunctions, etc. At the syntactic group level, for every identified group a syntactic head and a semantic head are selected.	
Texts	Media type	text	
	Linguality type	Monolingual	
	Languages	Polish	
		Language ID	PL
		Language script	UTF-8
	Modality	Modality type	Written language
	Size	1187 rules	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	UTF-8
Modality	Modality type	Written language
Size	1187 rules	

4.24. PANTERA

General Information

Short name	PANTERA
Description	Pantera is a morphosyntactic tagger based on Brill's Algorithm adapted for morphologically rich languages, e.g. Polish.
Identifier	424
Resource type	Tool/service
Tool/service type	Tool
URL	http://clip.ipipan.waw.pl/PANTERA/
Version	0.9-r150-4
Last update	2012-07-09

Contacts

Bartosz Zaborowski	
Contact	Jana Kazimierza 5 01-248 Warsaw bz233728@students.mimuw.edu.pl
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Szymon Acedański	
Contact	Jana Kazimierza 5 01-248 Warsaw accek@mimuw.edu.pl
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – unrestricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

GPL		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/PANTERA/	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2012-07-17	
Metadata creators	Michał Lenart	
	Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@ipipan.waw.pl http://zil.ipipan.waw.pl/MichalLenart
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Maciej Ogrodniczuk	
	Position	Assistant Professor

	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-17	

Usage

Foreseen use	NLP applications	
NLP-specific use	Morphosyntactic tagging Pos tagging	
Actual uses	NLP applications	
	NLP-specific use	Morphosyntactic tagging Pos tagging
	Reports	Przepiórkowski A., Bańko M., Górski R. L., Lewandowska-Tomaszczyk B., editors. Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]. PWN Scientific Publishers, Warsaw, 2012.
	Usage project	National Corpus of Polish
		Project short name NKJP
		URL http://www.nkjp.pl
		Funding type National funds
		Funder The Polish Ministry of Science and Higher Education (100%)
		Country Poland
		Start date 2007-12-13
		End date 2011-06-12
	Actual use details	Tagging of the National Corpus of Polish.
	NLP applications	
	NLP-specific use	Morphosyntactic tagging Pos tagging
	Reports	Ogrodniczuk M., Przepiórkowski A. Polish Language Processing

	Chains for Multilingual Information Systems. G. Bouma et al. (ed.): NLDB 2012, LNCS 7337, pp. 152–157. Springer, Heidelberg.	
Usage project	Applied Technology for Language-Aided CMS	
	Project short name	ATLAS
	URL	http://www.atlasproject.eu
	Funding type	EU funds National funds
	Funder	European Commission (50%) The Polish Ministry of Science and Higher Education (50%)
	Country	Poland
	Start date	2010-03-01
	End date	2013-02-28
Actual use details	Tagger for the Polish language processing chain.	
NLP applications		
NLP-specific use	Morphosyntactic tagging Pos tagging	
Reports	Ogrodniczuk M., Kopeć M. Rule-based coreference resolution module for Polish. In Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), pp. 191–200. Faro, Portugal.	
Usage project	Computer-based methods for coreference resolution in Polish texts	
	Project short name	CORE
	URL	http://zil.ipipan.waw.pl/CORE
	Funding type	National funds
	Funder	National Science Centre (100%)
	Country	Poland
	Start date	2011-04-18
	End date	2014-04-17
Actual use details	Tagger for the Polish coreference resolution module.	

Resource creation

Resource creator	Michał Lenart	
	Contact	Jana Kazimierza 5 01-248 Warsaw

		michal.lenart@ipipan.waw.pl http://zil.ipipan.waw.pl/MichalLenart
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Bartosz Zaborowski	
	Contact	Jana Kazimierza 5 01-248 Warsaw bz233728@students.mimuw.edu.pl
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Szymon Acedański	
	Contact	Jana Kazimierza 5 01-248 Warsaw accek@mimuw.edu.pl
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Funding projects	National Corpus of Polish	
	Project short name	NKJP
	URL	http://nkjp.pl/
	Funding type	National funds
	Funder	Polish Ministry of Science and Higher Education
	Country	Poland
	Start date	2007-01-01
	End date	2011-06-30
Creation start date	2010-03-01	

Resource documentation

Reports	Acedański S. A Morphosyntactic Brill Tagger for Inflectional Languages. Advances in Natural Language Processing, 2010, pp. 3-14.
Tool documentation type	Other

Tool/service

Tool/service type	Tool

Language dependent	False	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	Linux	
Required software	C++ Boost Libraries OpenMPI ICU4C library Morfusz (libmorfusz.so.0.6 or later, morphological analyzer for Polish) Java (JDK for source version) CMake (for source version) Autotools (for source version)	
Tool/service creation	Implementation language	C++
	Formalism	Brill tagger

4.25. PolNet – Polish Wordnet v.1

General Information

Short name	PolNet
Description	PolNet is a WordNet like lexical data base built from scratch according to the "merge model" methodology. Its design started in 2006 and continues. The resource development procedure is based on the exploration of good traditional dictionaries of Polish and language corpora investigations (IPI PAN Corpus and domain/application oriented corpora). The PolNet development was organized in an incremental way, starting with general and frequently used vocabulary. We selected the most frequent words found in a reference corpus of Polish language with however one important exception made for methodological reasons. The reason was that we assumed possibly early validation of the resource in a real-size application for which an application-complete vocabulary was necessary.
Identifier	425
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Wordnet
URL	http://lrc.amu.edu.pl/polnet/index.php
Version	1.0
Last update	2012-07-26

Contacts

--

Zygmunt Vetulani	
Position	Head of the Dept. of Computer Linguistics and Artificial Intelligence, Adam Mickiewicz University
Contact	Umultowska 87 61-614 Poznań vetulani@amu.edu.pl http://www.staff.amu.edu.pl/~vetulani/vetula_e.htm
Organization	Adam Mickiewicz University Department of Computer Linguistics and Artificial Intelligence vetulani@amu.edu.pl

Distribution

Availability	Available – restricted use	
IPR holder	Adam Mickiewicz University	
	Contact	Umultowska 87 61-614 Poznań vetulani@amu.edu.pl http://international.amu.edu.pl/
Availability start date	2011-11-25	

Licences

CC-BY-NC-ND		
Restrictions of use	Academic – non-commercial use Attribution No derivatives	
Access medium	Downloadable	
Download location	http://tc.amu.edu.pl/polnet/index.php	
Fee	free of charge	
Signatories	Zygmunt Vetulani	
	Contact	Umultowska 87 61-614 Poznań vetulani@amu.edu.pl http://www.staff.amu.edu.pl/~vetulani/vetula_e.htm
Distribution rights holder	Adam Mickiewicz University	
	Contact	Umultowska 87 61-614 Poznań vetulani@amu.edu.pl http://international.amu.edu.pl/

Metadata

Creation date	2012-07-26	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-26	

Usage

Foreseen use	Human use NLP applications
NLP-specific use	Document classification Semantic role labelling

Resource creation

Resource creator	Zygmunt Vetulani	
	Position	Head of the Dept. of Computer Linguistics and Artificial Intelligence, Adam Mickiewicz University
	Contact	Umultowska 87 61-614 Poznań vetulani@amu.edu.pl http://www.staff.amu.edu.pl/~vetulani/vetula_e.htm
	Organization	Adam Mickiewicz University Department of Computer Linguistics and Artificial Intelligence vetulani@amu.edu.pl
Creation start date	2006-01-01	

Resource documentation

Reports	Vetulani, Z., Walkowska, J., Obrębski, T., Konieczka, P., Rzepecki P., Marciniak, J. (2007): PolNet - Polish WordNet project algorithm, in: Z. Vetulani (ed.) Proceedings of the 3rd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, October 5-7, 2007, Poznań, Poland, Wyd. Poznańskie, Poznań, pp. 172-176. Pala, K., Horák, A., Rambousek, A., Vetulani, Z., Konieczka, P., Marciniak, J.,
----------------	---

	<p>Obrębski, T., Rzepecki P., Walkowska, J., (2007): DEB Platform tools for effective development of WordNets in application to PolNet, in: Z. Vetulani (ed.) Proceedings of the 3rd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, October 5-7, 2007, Poznań, Poland, Wyd. Poznańskie, Poznań, pp. 514-518.</p> <p>Vetulani, Z., Walkowska, J., Obrębski, T., Marciniak, J., Konieczka, P., Rzepecki, P. (2009): An Algorithm for Building Lexical Semantic Network and Its Application to PolNet - Polish WordNet Project, in: Vetulani, Z. and Uszkoreit, H., Human Language Technology. Challenges of the Information Society. LTC 2007. Revised selected papers. LNAI 5603, Springer, 369-381.</p> <p>Vetulani, Z., Obrębski, T. (2010): Resources for Extending the PolNet-Polish WordNet with a Verbal Component, in: Bhattacharyya, P., Fellbaum, Ch., Vossen, P. (eds.) Principles, Construction and Application of Multilingual Wordnets. Proceedings of the 5th Global Wordnet Conference, Narosa Publishing House: New Delhi, Chennai, Mumbai, Kolkata, pp. 325-330.</p> <p>Vetulani, Z., Kubis, M., Obrębski, T. (2010): PolNet – Polish WordNet: Data and Tools, in: Calzolari, N. (ed.) Proceedings of the seventh International conference on Language Resources and Evaluation (LREC 2010), May 19-21, Valletta, Malta, (Proceedings), ELRA, Paris. http://www.lrec-conf.org/proceedings/lrec2010/summaries/947.html</p> <p>Vetulani, Z., Marcinak, J., Obrębski, T., Vetulani, G., Dabrowski, A., Kubis, M., Osiński, J., Walkowska, J., Kubacki, P., Witalewski, K. (2010): Zasoby językowe i technologie przetwarzania tekstu. POLINT-112-SMS jako przykład aplikacji z zakresu bezpieczeństwa publicznego (in Polish) (Language resources and text processing technologies. POLINT-112-SMS as example of homeland security oriented application), ISBN 978-83-232-2155-5, ISSN 1896-379X, Adam Mickiewicz University Press: Poznań.</p> <p>Vetulani, Z. (2012): Wordnet Based Lexicon Grammar for Polish, in: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk and Stelios Piperidis (eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), May 23-25, 2012, Istanbul, Turkey, ELRA, Paris, isbn: 978-2-9517408-7-7 (accessible through http://www.lrec-conf.org/proceedings/lrec2012/index.html)</p> <p>Vetulani, Z. (2012): Language Resources in a Public Security Application with Text Understanding Competence. A Case Study: POLINT-112-SMS, in: Vetulani, Z., Geoffrois, E. (eds.), Proceedings of LREC 2012 Workshop on Language Resources for Public Security Applications, May 27, 2012, Istanbul, ELRA, Paris, pp. 54-63. (http://www.lrec-conf.org/proceedings/lrec2012/index.html).</p>
--	---

Lexical conceptual resource

Lexical conceptual resource type	Wordnet	
Lexical conceptual resource encoding	Encoding level	Semantics
	Linguistic information	Part of speech Semantics – cross references Semantics – relations Semantics – relations – antonyms Semantics – relations – hyperonyms Semantics – relations – hyponyms Semantics – relations – meronyms

		Semantics – relations – synonyms
	Conformance to standards best practices	Word net

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	Latin
Modality	Modality type	Written language
Size	8.0 mb, 13200 synsets	

4.26. Polish Wikipedia Corpus

General Information

Short name	PWC
Description	Full textual content of Polish Wikipedia (http://pl.wikipedia.org/) on 28.04.2012. The main directory contains subdirectories named AA to DQ. Each of those consists of subdirectories 00 to 99, containing approximately 100 kB of text each, one Wikipedia article per file. Article files start with a title, followed by a blank line. Only ordinary articles are present - without stubs, templates, disambiguation pages, history of changes etc. All the multimedia, tables, references, links, and other non-plaintext elements have been removed. Text is encoded as UTF-8. In 839 269 articles there are 127 million segments, 918 MB of text in total. The corpus has been created by applying WikiExtractor script (http://medialab.di.unipi.it/wiki/Wikipedia_Extractor) to Wikipedia dump (http://dumps.wikimedia.org/backup-index.html) and dividing 100 kB files into individual articles using own code. Wikipedia's text content is released under the Creative Commons Attribution-Share-Alike License 3.0 (http://creativecommons.org/licenses/by-sa/3.0/).
Identifier	426
Resource type	Corpus
URL	http://clip.ipipan.waw.pl/PolishWikipediaCorpus
Version	1.0
Last update	2012-10-03

Contacts

Piotr Przybyła	
Position	PhD student

Contact	Jana Kazimierza 5 01-248 Warsaw P.Przybyla@phd.ipipan.waw.pl
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

CC-BY-SA		
Restrictions of use	Attribution	
	Share alike	
Access medium	Downloadable	
Download location	http://clip.ipipan.waw.pl/PolishWikipediaCorpus	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2012-10-03	
Metadata creators	Piotr Przybyla	
	Position	PhD student
	Contact	Jana Kazimierza 5 01-248 Warsaw P.Przybyla@phd.ipipan.waw.pl
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Source	CESAR	
Metadata language	en	

ID	
Metadata last date updated	2012-10-03

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
Size	839269 articles	
Text classification	Text type	encyclopaedic
Creation	Original source	Polish Wikipedia dump http://dumps.wikimedia.org/backup-index.html
	Creation mode	Automatic
	Creation mode details	Applied WikiExtractor script to Wikipedia dump (http://dumps.wikimedia.org/backup-index.html) and divided 100 kB files into individual articles using own code.
	Creation tools	WikiExtractor script (http://medialab.di.unipi.it/wiki/Wikipedia_Extractor)

4.27. SEJFEK4Spejd

General Information

Short name	SEJFEK4Spejd
Description	SEJFEK4Spejd is the SEJFEK lexicon (Grammatical Lexicon of Polish Economical Phraseology) converted into a lexicalized Spejd shallow grammar. It contains 11,270 automatically generated rules which recognize inflected, case-insensitive multi-word economic terms from the lexicon. Recognized multi-word terms are combined into syntactic words. During the analysis disambiguation (unification and POS-based selection of interpretations) of terms is also performed.
Identifier	427
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Computational lexicon
URL	http://zil.ipipan.waw.pl/SEJFEK4Spejd
Version	0.2
Last update	2011-10-11

Contacts

Bartosz Zaborowski	
---------------------------	--

Contact	Jana Kazimierza 5 01-237 Warsaw bartosz.zaborowski@ipipan.waw.pl
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Aleksandra Wieczorek	
Contact	Jana Kazimierza 5 01-237 Warsaw aleksandra.wieczorek@ipipan.waw.pl
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – restricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-237 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

GPL		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/SEJFEK4Spejd?action=AttachFile&do=get&target=SEJFEK4Spejd_converter.tar.gz	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-237 Warsaw

		ipi@ipipan.waw.pl http://www.ipipan.eu
CC-BY-SA		
Restrictions of use	Attribution Share alike	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/SEJFEK4Spejd?action=AttachFile&do=get&target=SEJFEK4Spejd_source_lexicon.tar.gz http://zil.ipipan.waw.pl/SEJFEK4Spejd?action=AttachFile&do=get&target=SEJFEK4Spejd.tar.gz	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-237 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2013-01-22	
Metadata creators	Bartosz Zaborowski	
	Contact	Jana Kazimierza 5 01-237 Warsaw bartosz.zaborowski@ipipan.waw.pl
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Source	Nekst	
Metadata language ID	en	
Metadata last date updated	2013-01-23	

Resource creation

Resource creator	Bartosz Zaborowski	
	Contact	Jana Kazimierza 5 01-237 Warsaw

	bartosz.zaborowski@ipipan.waw.pl
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Aleksandra Wieczorek	
Contact	Jana Kazimierza 5 01-237 Warsaw aleksandra.wieczorek@ipipan.waw.pl
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Funding projects	NEKST — Adaptive system supporting solving problems on the basis of content analysis of electronic documents
	Project ID POIG.01.01.02-14-013/09
	Funding type EU funds National funds
	Country Poland
	Start date 2009-04-01
	End date 2014-02-10
Creation start date	2011-10-01

Resource documentation

Reports	SAVARY, A., ZABOROWSKI, B., KRAWCZYK-WIECZOREK, A., MAKOWIECKI, F. (2012): SEJFEK — a Lexicon and a Shallow Grammar of Polish Economic Multi- Word Units, in Proceedings of Cognitive Aspects of the Lexicon (COGALEX-III), a Workshop at COLING 2012, Mumbai, India.
----------------	---

Lexical conceptual resource

Lexical conceptual resource type	Computational lexicon	
Lexical conceptual resource encoding	Encoding level	Morphology Syntax
	Linguistic information	Lemma – multi word units Case Degree Gender Number Person Part of speech

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	Latin
Modality	Modality type	Other
Size	11270 rules	

4.28. PNET

General Information

Short name	PNET
Description	PNET (Polish Named Entity Triggers) is an electronic lexicon containing partly inflected external or internal evidences, or trigger words, for Polish named entities (NEs). A NE trigger is a word or a list of words which appears frequently in the vicinity or inside named entities and is a good indicator of these NEs' types. For instance "aktor" ('actor') is an external evidence for person names (as in "aktor [Zbigniew Buczkowski]"), while "von" is an internal evidence for the same type ("[John von Neumann]"). Many words can be both external and internal evidences, e.g. "jezioro" ('lake') is a external evidence in "jezioro [Mamry]" ('[Mamry] lake') and an internal evidence in "Jezioro Białe" ('[White Lake]'). External and internal NE evidences can be used in automatic NE recognition via grammar-based or machine-learning methods.
Identifier	428
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://zil.ipipan.waw.pl/PNET
Version	1.0
Last update	2012-10-25

Contacts

Agata Savary	
Position	Associate Professor
Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
Organization	Université François Rabelais Tours Laboratoire d'Informatique

	li@univ-tours.fr
--	--

Distribution

Availability	Available – restricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu
Availability start date	2012-10-30	

Licences

BSD-style		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/PNET?action=AttachFile&do=get&target=PNET.tar.gz	
Fee	free of charge	
Signatories	Agata Savary	
	Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2012-10-25
Metadata creators	Agata Savary

	Position	Associate Professor
	Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-10-25	

Usage

Foreseen use	NLP applications
NLP-specific use	Named entity recognition

Resource creation

Resource creator	Agata Savary	
	Position	Associate Professor
	Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
	Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr
	Malgorzata Baron	
	Position	linguist
	Contact	gggossiaaa@gmail.com
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Leszek Manicki	
	Position	programmer
	Contact	lebiega@gmail.com
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Creation start date	2011-10-20	

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology Semantics
	Linguistic information	Lemma Case Gender Number Other Part of speech Semantics – semantic class
Creation	Original source	http://pl.wikipedia.org http://zil.ipipan.waw.pl/PoliMorf http://www.cnrtl.fr/lexiques/prolex/
	Creation mode	Mixed
	Creation mode details	See README.txt file in the downloadable archive.

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	Latin
Modality	Modality type	Written language
Size	1503 entries, 28085 words	

4.29. An LFG grammar of Polish (POLFIE)

General Information

Short name	POLFIE
Description	POLFIE is an LFG grammar of Polish implemented in the XLE system (Xerox Linguistic Environment), it is being developed within NEKST project. It provides a two-layer representation: constituent structure (c-structure, tree representation) and functional structure (f-structure, AVM representation). It is based on two previous implemented grammars of Polish: its c-structure is based on GFJP2, a DCG grammar used by the parser Świgr, while its f-structure is inspired by FOJP, an HPSG grammar of Polish. Lexical entries used by the grammar are created using Morfeusz, the state-of-the-art morphological analyser for Polish, and converted valence dictionaries used by Świgr.
Identifier	429

Resource type	Language description
Language description type	Grammar
URL	http://zil.ipipan.waw.pl/LFG
Version	1.0
Last update	2013-01-23

Contacts

Agnieszka Patejuk	
Position	grammar author
Contact	Jana Kazimierza 5 01-248 Warsaw aep@ipipan.waw.pl

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

GPL	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/LFG
Fee	free of charge

Metadata

Creation date	2013-01-23	
Metadata creators	Agnieszka Patejuk	
	Contact	Jana Kazimierza 5 01-248 Warsaw aep@ipipan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-23	

Usage

Foreseen use	NLP applications
NLP-specific use	Parsing

Resource creation

Resource creator	Agnieszka Patejuk	
	Contact	Jana Kazimierza 5 01-248 Warsaw aep@ipipan.waw.pl
Funding projects	An adaptive system to support problem-solving on the basis of document collections in the Internet	
	Project short name	NEKST
	URL	http://www.ipipan.waw.pl/nekst/
	Funding type	National funds
	Funder	The Polish Ministry of Science and Higher Education
	Country	Poland
	Start date	2009-04-01
	End date	2014-02-14
Creation start date	2010-03-09	

Language description

Language description type	Grammar	
Language description encoding	Encoding level	Syntax
	Theoretic model	LFG (Lexical Functional Grammar) grammars minimally provide two levels of representation: constituent structure (c-structure) produced by context-free phrase structure rules and functional structure (f-structure) created by functional descriptions.
Texts	Media type	text
	Linguality type	Monolingual
	Languages	Polish
		Language ID PL
		Language script UTF-8
	Modality	Modality type Written language
	Size	Grammar has 46 rules with 355 states, 634 arcs, and 1100 disjuncts

	(2317 DNF).
--	-------------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	UTF-8
Modality	Modality type	Written language
Size	Grammar has 46 rules with 355 states, 634 arcs, and 1100 disjuncts (2317 DNF).	

4.30. Lexeme Forge

General Information

Short name	LF
Description	Lexeme Forge is a web application designed to allow collaborative creating of inflectional dictionaries.
Identifier	430
Resource type	Tool/service
Tool/service type	Nlp development environment
URL	http://zil.ipipan.waw.pl/Kuźnia
Version	0.9
Last update	2013-01-14

Contacts

Jan Szejko	
Contact	Jana Kazimierza 5 01-248 Warsaw jan.szejko@ipipan.waw.pl

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

BSD-style	
------------------	--

Restrictions of use	Attribution
Access medium	Downloadable
Download location	https://bitbucket.org/janek37/lexeme_forge/downloads/lexeme_forge-0.9.tar.gz
Fee	free of charge

Metadata

Creation date	2013-01-14	
Metadata creators	Jan Szejko	
	Contact	Jana Kazimierza 5 01-248 Warsaw jan.szejko@ipipan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-14	

Usage

Foreseen use	NLP applications
NLP-specific use	Lexicon enhancement
Foreseen use	NLP applications
NLP-specific use	Lexicon merging

Resource creation

Resource creator	Jan Szejko	
	Position	Programmer
	Contact	Jana Kazimierza 5 01-248 Warsaw jan.szejko@ipipan.waw.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%)

		Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2011-03-30	

Tool/service

Tool/service type	Nlp development environment	
Language dependent	True	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	Linux	
Required software	Python (version 2.6 or higher) Django (version 1.3 or higher) PostgreSQL (version 9.1 or higher)	
Tool/service creation	Implementation language	Python JavaScript SQL

4.31. A tool for creating a Polish Valence Dictionary

General Information

Short name	Slowal
Description	Slowal is a web tool designed for creating valence dictionaries based on the format presented by Filip Skwarski. It describes each lemma by a list of individual frames presented as tables which can be expanded by adding new positions, arguments, series of characteristics and examples showing usage of the frame in the Polish language.
Identifier	431
Resource type	Tool/service
Tool/service type	Nlp development environment
URL	http://zil.ipipan.waw.pl/Slowal
Version	1.0
Last update	2013-01-18

Contacts

Bartłomiej Nitoń	
Contact	Jana Kazimierza 5 01-248 Warsaw bartek.niton@gmail.com
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

BSD-style	
Restrictions of use	Attribution
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/Slowal?action=AttachFile&do=get&target=Download
Fee	free of charge

Metadata

Creation date	2013-01-18	
Metadata creators	Bartłomiej Nitoń	
	Contact	Jana Kazimierza 5 01-248 Warsaw bartek.niton@gmail.com
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-18	

Resource creation

Resource creator	Nitoń Bartłomiej	
	Position	Programmer

	Contact	Jana Kazimierza 5 01-248 Warsaw bartek.niton@gmail.com
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Filip Skwarski	
	Position	Linguist
	Contact	Jana Kazimierza 5 01-248 Warsaw fskwarski@gmail.com
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2010-03-01	

Tool/service

Tool/service type	Nlp development environment	
Language dependent	True	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	Linux	

Required software	Django (version 1.3 or higher) PostgreSQL (version 9.1 or higher) Python (version 2.7 or higher)	
Tool/service creation	Implementation language	Python JavaScript

4.32. CorpCor

General Information

Short name	CorpCor
Description	CorpCor is a web-based tool for correcting morphosyntactic annotation in TEI XML encoded corpora (e.g. National Corpus of Polish).
Identifier	432
Resource type	Tool/service
Tool/service type	Tool
URL	http://zil.ipipan.waw.pl/CorpCor
Version	1.1
Last update	2012-09-24

Contacts

Łukasz Kobyliński	
Position	Research Assistant
Contact	Jana Kazimierza 5 01-248 Warsaw lkobylnski@ipipan.waw.pl http://zil.ipipan.waw.pl/LukaszKobylnski

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

GPL	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/CorpCor?action=AttachFile&do=get&target=CorpCor-1.1.zip
Fee	free of charge

Metadata

Creation date	2013-01-10	
Metadata creators	Łukasz Kobyliński	
	Position	Research Assistant
	Contact	Jana Kazimierza 5 01-248 Warsaw lkobyliński@ipipan.waw.pl http://zil.ipipan.waw.pl/LukaszKobyliński
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-10	

Usage

Foreseen use	NLP applications	
NLP-specific use	Pos tagging	
Actual uses	NLP applications	
	NLP-specific use	Pos tagging
	Usage project	Automatic detection and correction of annotation errors in Polish language corpora
		URL http://zil.ipipan.waw.pl/Automatic detection and correction of annotation errors in Polish language corpora
		Funding type National funds
		Funder The Polish Ministry of Science and Higher Education (100%)
		Country Poland
		Start date 2012-01-01
		End date 2013-12-30
	Actual use details	Manual correction of pos-tagging errors in the manually annotated subcorpus of National Corpus of Polish.

Resource creation

Resource creator	Łukasz Kobyliński	
	Position	Research Assistant

	Contact	Jana Kazimierza 5 01-248 Warsaw lkobylnski@ipipan.waw.pl http://zil.ipipan.waw.pl/LukaszKobylnski
Funding projects	Automatic detection and correction of annotation errors in Polish language corpora	
	URL	http://zil.ipipan.waw.pl/Automatic detection and correction of annotation errors in Polish language corpora
	Funding type	National funds
	Funder	The Polish Ministry of Science and Higher Education (100%)
	Country	Poland
	Start date	2012-01-01
	End date	2013-12-30
Creation start date	2012-01-01	

Resource documentation

Reports	Included README file.
Tool documentation type	Help functions

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	Linux Windows	
Required software	Java web application server (e.g. Tomcat)	
Tool/service creation	Implementation language	Java

4.33. plWikiEcono

General Information

Short name	plWikiEcono
Description	A corpus of Polish Wikipedia articles from the domain of economy. Automatically

	annotated using TaKIPi 1.8, TEI format.
Identifier	433
Resource type	Corpus
URL	http://zil.ipipan.waw.pl/plWikiEcono
Version	1.0

Contacts

Łukasz Kobyliński	
Position	Research Assistant
Contact	Jana Kazimierza 5 01-248 Warsaw lkobyliński@ipipan.waw.pl http://zil.ipipan.waw.pl/LukaszKobyliński

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

CC-BY-SA	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/plWikiEcono?action=AttachFile&do=get&target=wikipedia-econotei.7z
Fee	free of charge

Metadata

Creation date	2013-01-10	
Metadata creators	Łukasz Kobyliński	
	Position	Research Assistant
	Contact	Jana Kazimierza 5 01-248 Warsaw lkobyliński@ipipan.waw.pl http://zil.ipipan.waw.pl/LukaszKobyliński
Source	CESAR	
Metadata language ID	en	
Metadata last date	2013-01-10	

updated	
---------	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
Size	34.0 mb, 933892 words	
Annotation	Segmentation	
	Annotation standoff	True
	Segmentation level	Paragraph
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode details	inherited from source corpus (no need to generate new segmentation when sampling)
	Start date	2011-04-01
	End date	2011-04-30
	Segmentation	
	Annotation standoff	True
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Automatic
	Annotation tool	TaKIPI 1.8
	Start date	2011-04-01
	End date	2011-04-30
	Segmentation	
	Annotation standoff	True
	Segmentation level	Word

Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation tool	TaKIPI 1.8
Start date	2011-04-01
End date	2011-04-30
Lemmatization	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation tool	TaKIPI 1.8
Start date	2011-04-01
End date	2011-04-30
Morphosyntactic annotation - below POS tagging	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation tool	TaKIPI 1.8
Start date	2011-04-01
End date	2011-04-30
Morphosyntactic annotation – POS tagging	

	Annotation standoff	True
	Segmentation level	Word
	Format	text/xml
	Tagset	NKJP tagset
	Conformance to standards best practices	TEI
	Annotation mode	Automatic
	Annotation tool	TaKIPI 1.8
	Start date	2011-04-01
	End date	2011-04-30
Creation	Original source	Polish Wikipedia
	Creation mode	Mixed
	Creation mode details	Economy-related categories from the Polish Wikipedia, including economy-related subcategories, stripped Wikipedia annotations, tagged with TaKIPI 1.8 and converted to TEI format.
	Creation tools	TaKIPI 1.8 Java code

4.34. plWikiEconoSenses

General Information

Short name	plWikiEconoSenses
Description	A corpus of Polish Wikipedia articles from the domain of economy. Annotated automatically (morphosyntatic layer) and manually (word sense layer), TEI format.
Identifier	434
Resource type	Corpus
URL	http://zil.ipipan.waw.pl/plWikiEcono
Version	1.0

Contacts

Lukasz Kobyliński	
Position	Research Assistant
Contact	Jana Kazimierza 5 01-248 Warsaw

	lkobylnski@ipipan.waw.pl http://zil.ipipan.waw.pl/LukaszKobylnski
--	---

Distribution

Availability	Available – unrestricted use
--------------	------------------------------

Licences

CC-BY-SA	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/plWikiEcono?action=AttachFile&do=get&target=wikipedia-econotei-senses.7z
Fee	free of charge

Metadata

Creation date	2013-01-10	
Metadata creators	Lukasz Kobyliński	
	Position	Research Assistant
	Contact	Jana Kazimierza 5 01-248 Warsaw lkobylnski@ipipan.waw.pl http://zil.ipipan.waw.pl/LukaszKobylnski
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-10	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
Size	16.5 mb, 438865 words	
Annotation	Segmentation	
	Annotation standoff	True
	Segmentation	Paragraph

level	
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode details	inherited from source corpus (no need to generate new segmentation when sampling)
Start date	2011-04-01
End date	2011-04-30
Segmentation	
Annotation standoff	True
Segmentation level	Sentence
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation tool	TaKIPI 1.8
Start date	2011-04-01
End date	2011-04-30
Segmentation	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation tool	TaKIPI 1.8
Start date	2011-04-01
End date	2011-04-30
Lemmatization	
Annotation	True

standoff	
Segmentation level	Word
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation tool	TaKIPi 1.8
Start date	2011-04-01
End date	2011-04-30
Morphosyntactic annotation - below POS tagging	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation tool	TaKIPi 1.8
Start date	2011-04-01
End date	2011-04-30
Morphosyntactic annotation – POS tagging	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation tool	TaKIPi 1.8

	Start date	2011-04-01
	End date	2011-04-30
	Semantic annotation – word senses	
	Annotation standoff	True
	Segmentation level	Word
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Manual
	Annotation mode details	manually disambiguated using AnotEk
	Annotation tool	AnotEk
	Start date	2011-04-01
	End date	2011-11-19
Creation	Original source	Polish Wikipedia
	Creation mode	Mixed
	Creation mode details	Economy-related categories from the Polish Wikipedia, including economy-related subcategories, stripped Wikipedia annotations, tagged with TaKIPI 1.8 and converted to TEI format.
	Creation tools	TaKIPI 1.8 Java code AnotEk 1.0

4.35. Prolexbase

General Information

Short name	Prolexbase
Description	Prolexbase is a multilingual relational dictionary of proper names, conceived at the University of Tours, France and further developed at the University of Belgrade, Serbia, and at the Polish Academy of Sciences (IPIPAN). It contains a language-independent typology of proper names with 4 supertypes and 34 types, as well as various language-independent or language-specific relations (synonymy, meronymy accessibility, variation etc.). A pivot-oriented design of concepts yields alignment of proper names in a language with their counterparts in other languages. Currently, the resources counts about 40,000 Polish, 33,000 English and 100,000 French proper names together with their corresponding 165,000 Polish, 18,000 English and 142,393 English inflected forms. A large majority of the data have been extracted from Wikipedia and GeoNames. All data have been manually validated.

Identifier	435
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://zil.ipipan.waw.pl/Prolexbase
Version	2.0
Last update	2012-01-22

Contacts

Agata Savary	
Position	Associate Professor
Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr

Distribution

Availability	Available – restricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu
	Université François Rabelais Tours	
	Short name	UFRT
	Department name	Laboratoire d'Informatique
	Contact	64 avenue Jean Portalis 37200 Tours li@univ-tours.fr http://li.univ-tours.fr

Licences

CC-BY-SA		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/Prolexbase?action=AttachFile&do=get&target=prolexbase.mysql.tar.gz	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu
	Laboratoire d'informatique	
	Short name	LI
	Department name	Databases and Natural Language Processing Group (BdThn)
	Contact	64 avenue Jean Portalis 37200 Tours li@univ-tours.fr http://li.univ-tours.fr

Metadata

Creation date	2013-01-22	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Agata Savary	
	Position	Associate Professor
	Contact	3, place Jean-Jaurès 41000 Blois

		agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
	Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-22	

Usage

Foreseen use	NLP applications
NLP-specific use	Coreference resolution Information extraction Knowledge representation Machine translation Morphological analysis Named entity recognition Question answering Semantic web

Resource creation

Resource creator	Agata Savary	
	Position	Associate Professor
	Contact	3, place Jean-Jaurès 41000 Blois agata.savary@univ-tours.fr http://www.info.univ-tours.fr/~savary
	Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr
	Malgorzata Baron	
	Position	Linguist
	Contact	gggossiaaa@gmail.com
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Béatrice Bouchou-Markhoff	
	Position	Associate Professor

	Contact	3, place Jean-Jaurès 41000 Blois beatrice.bouchou@univ-tours.fr http://www.info.univ-tours.fr/~bouchou/index_a.html
	Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr
	Leszek Manicki	
	Position	Analyst and Programmer
	Contact	lebiega@gmail.com
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Denis Maurel	
	Position	Professor
	Contact	64, Avenue Jean Portalis 37200 Tours denis.maurel@univ-tours.fr http://www.univ-tours.fr/acces-rapide/m-maurel-denis-84407.kjsp
	Organization	Université François Rabelais Tours Laboratoire d'Informatique li@univ-tours.fr
	Duško Vitas	
	Position	Professor
	Contact	Studentski trg 16 11000 Belgrade vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas/index-en.html
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Funding projects	NEKST — Adaptive system supporting solving problems on the basis of content analysis of electronic documents	
	Project ID	POIG.01.01.02-14-013/09
	Funding type	EU funds National funds
	Country	Poland
	Start date	2009-04-01
	End date	2014-02-10
Creation start date	2011-01-03	

Resource documentation

Reports	Savary, A., Manicki, L., Baron, M.: ProlexFeeder— Populating a Multilingual Ontology of Proper Names from Open Sources. Submitted to Journal of Language Modelling. Bouchou, B., Maurel, D. (2008): Prolexbase et LMF : vers un standard pour les ressources lexicales sur les noms propres. In Traitement Automatique des Langues, 49(1). Krstev S., Vitas D., Maurel D., Tran M. (2005), Multilingual Ontology of Proper Names, LTC'05, Poznan, Poland.
----------------	---

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology
	Linguistic information	<p>Lemma</p> <p>Lemma – abbreviations</p> <p>Lemma – multi word units</p> <p>Case</p> <p>Morpho – derivation</p> <p>Gender</p> <p>Inflection</p> <p>Other</p> <p>Part of speech</p> <p>Semantics – cross references</p> <p>Semantics – relations</p> <p>Semantics – relations – hyperonyms</p> <p>Semantics – relations – hyponyms</p> <p>Semantics – relations – meronyms</p> <p>Semantics – relations – synonyms</p> <p>Semantics – traits</p> <p>Translation equivalent</p> <p>Usage – frequency</p> <p>Usage – register</p>
Creation	Original source	http://www.cnrtl.fr/lexiques/prolex/
	Creation mode	Mixed
	Creation mode details	see "ProlexFeeder— Populating a Multilingual Ontology of Proper Names from Open Sources"
	Creation tools	<p>ProlexFeeder</p> <p>Translatca's (http://www.translatca.pl/) automatic inflection tool for multi-word units</p>

Texts

Media type	text

Linguality type	Multilingual	
Languages	Polish	
	Language ID	PL
	Language script	latin
	English	
	Language ID	EN
	Language script	latin
	French	
	Language ID	FR
	Language script	latin
Size	67,000 concepts, 173,000 entries, 325,000 units, 65,500	
Character encoding	UTF-8	

4.36. Dependency Parsing Model for Polish

General Information

Short name	Polish Dependency Parsing Model
Description	Statistical dependency parsing model is trained on the Polish Dependency Bank (PDB, Pol. Składnica zależnościowa) with the publicly available parsing system -- MaltParser. MaltParser is a transition-based dependency parser that uses a deterministic parsing algorithm. The deterministic parsing algorithm builds a dependency structure of an input sentence based on transitions (shift-reduce actions) predicted by a classifier. The classifier learns to predict the next transition given training data and the parse history.
Identifier	436
Resource type	Tool/service
Tool/service type	Tool
URL	http://zil.ipipan.waw.pl/PolishDependencyParser
Version	0.1
Last update	2013-01-16

Contacts

Alina Wróblewska	
Contact	Jana Kazimierza 5 01-248 Warsaw alina@ipipan.waw.pl

Distribution

Availability	Available – restricted use
--------------	----------------------------

Licences

GPL	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/PolishDependencyParser
Fee	free of charge

Metadata

Creation date	2013-01-16	
Metadata creators	Alina Wróblewska	
	Position	Assistant
	Contact	Jana Kazimierza 5 01-248 Warsaw alina@ipipan.waw.pl http://zil.ipipan.waw.pl/AlinaWroblewska
Source	NEKST	
Metadata language ID	en	
Metadata last date updated	2013-01-16	

Usage

Foreseen use	NLP applications	
NLP-specific use	Parsing	
Actual uses	NLP applications	
	NLP-specific use	Parsing
	Reports	Alina Wróblewska. 2012. Polish dependency bank. Linguistic Issues in Language Technology, 7(1). Alina Wróblewska and Adam Przepiórkowski. Induction of dependency structures based on weighted projection. In Proceedings of the 4th International Conference on Computational Collective Intelligence Technologies and Applications (ICCCI 2012), Part I, volume 7653 of Lecture Notes in Artificial Intelligence, pages 364–374, Berlin, 2012. Springer-Verlag.

Resource creation

Resource creator	Alina Wróblewska	
	Position	Computational linguist
	Contact	Jana Kazimierza 5 01-248 Warsaw alina@gmail.com http://zil.ipipan.waw.pl/AlinaWroblewska
Funding projects	An adaptive system to support problem-solving on the basis of document collections in the Internet	
	Project short name	NEKST
	URL	http://www.ipipan.waw.pl/nekst/
	Funding type	EU funds
	Funder	European Regional Development Fund
	Country	EU
	Start date	2009-04-01
	End date	2014-02-10
Creation start date	2013-01-16	

Resource documentation

Reports	Alina Wróblewska and Marcin Woliński. 2012. Preliminary experiments in Polish dependency parsing. In Pascal Bouvry, Mieczysław A. Kłopotek, Franck Leprevost, Małgorzata Marciniak, Agnieszka Mykowiecka, and Henryk Rybiński, editors, Security and Intelligent Information Systems: International Joint Conference, SIIS 2011, Warsaw, Poland, June 13-14, 2011, Revised Selected Papers, volume 7053 of Lecture Notes in Computer Science, pages 279–292. Springer-Verlag.
Tool documentation type	Help functions

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	Linux	

Required software	MaltParser(1.7.2 or higher)		
Tool/service evaluation	Evaluated	True	
	Level	Diagnostic	
	Type	Black box	
	Criteria	Intrinsic	
	Measure	Automatic	
	Reports	Evaluation results can be found in: Alina Wróblewska and Marcin Woliński. Preliminary experiments in Polish dependency parsing. In Security and Intelligent Information Systems: International Joint Conference, SIIS 2011, Warsaw, Poland, June 13-14, 2011, Revised Selected Papers, volume 7053 of Lecture Notes in Computer Science, pages 279–292. Springer-Verlag, 2012.	
	Details	Polish MaltParser trained on 6832 PDB-dependency structures is evaluated against a set of 759 sentences taken from PDB and a set of 50 manually annotated sentences taken from Polish magazines. Evaluation metrics: labelled attachment score (LAS) and unlabelled attachment score (UAS). Polish MaltParser achieves 84.7% LAS and 90.5% UAS if tested against the PDB validation set and 68.5% LAS/72.2% UAS if tested against the set of 50 manually annotated sentences.	
	Evaluators	Alina Wróblewska	
		Contact	Jana Kazimierza 5 01-248 Warsaw alina@ipipan.waw.pl

4.37. Polish HateSpeech Corpus

General Information

Short name	PL hatespeech
Description	HateSpeech corpus in the current version contains over 2000 posts crawled from public Polish web. They represent various types and degrees of offensive language, expressed toward minorities (eg. ethnical, racial). The data were annotated manually.
Identifier	437
Resource type	Corpus
URL	http://www.raportmniejszosci.pl
Version	2
Last update	2011-09-01

Contacts

Marek Troszyński	
Position	Principal

Contact	al. Waszyngtona 67/18 04-074 Warszawa marek.troszynski@wiedzalokalna.pl http://www.raportmniejszosci.pl
Organization	Fundacja Wiedza Lokalna jadwiga.janik@linkdodialogu.pl

Distribution

Availability	Available – unrestricted use	
IPR holder	Fundacja Wiedza Lokalna	
	Contact	al. Waszyngtona 67/18 04-074 Warsaw biuro@wiedzalokalna.pl http://www.wiedzalokalna.pl/

Licences

CC-BY-NC-SA		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/HateSpeech/hatespeech.tar.bz2	
Distribution rights holder	Fundacja Wiedza Lokalna	
	Contact	al. Waszyngtona 67/18 04-074 Warsaw biuro@wiedzalokalna.pl http://www.wiedzalokalna.pl

Metadata

Creation date	2012-06-16	
Metadata creators	Marek Troszyński	
	Position	Principal
	Contact	al. Waszyngtona 67/18 04-074 Warsaw marko.tadic@ffzg.hr
	Organization	Fundacja Wiedza Lokalna biuro@wiedzalokalna.pl
Metadata language ID	en	
Metadata last date	2011-11-27	

updated	
---------	--

Resource creation

Resource creator	Fundacja Wiedza Lokalna, Department/Institute of Linguistics	
	Contact	al. Waszyngtona 67/18 04-074 Warsaw biuro@wiedzalokalna.pl http://www.wiedzalokalna.pl/
Funding projects	Demokracja w Działaniu	
	URL	http://www.batory.org.pl/programy_dotacyjne/demokracja_w_dzialaniu
	Funding type	Own funds
	Funder	Fundacja Batorego (100%)
	Start date	2010-02-01
	End date	2012-06-30
Creation start date	2010-03-01	

Resource documentation

Reports	In preparation
---------	----------------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	pl
	Language script	latin
Size	1.2 mb, 2000 units	
Character encoding	UTF-8	
Annotation	Alignment	
	Segmentation level	Paragraph

4.38. Polish Coreference Corpus

General Information

Description	The Polish Coreference Corpus (PL: Polski Korpus Koreferencyjny) is a result of the "Computer-based methods for coreference resolution in Polish texts" project. It contains
-------------	--

	short fragments (250-350 segments each) of texts randomly selected (preserving the original text type balance) from the full version of the National Corpus of Polish. These fragments are manually annotated with identity coreferential chains and quasi-identity relations. The corpus is supplied in two xml-based formats: MMAX and TEI. It contains automatic morphosyntactic annotation, in TEI format it also has automatic named entity and shallow parsing annotations.
Identifier	438
Resource type	Corpus
URL	http://zil.ipipan.waw.pl/PolishCoreferenceCorpus
Version	0.5
Last update	2013-01-08

Contacts

Maciej Ogrodniczuk	
Position	Assistant Professor
Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Mateusz Kopec	
Contact	m.kopec@ipipan.waw.pl http://zil.ipipan.waw.pl/MateuszKopec
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – unrestricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

CC-BY	
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/PolishCoreferenceCorpus
Fee	free of charge

Metadata

Creation date	2013-01-08
Metadata creators	Mateusz Kopec
	Contact m.kopec@ipipan.waw.pl http://zil.ipipan.waw.pl/MateuszKopec
	Organization Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Metadata last date updated	2013-01-22

Resource creation

Resource creator	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu
Funding projects	Computer-based methods for coreference resolution in Polish texts	
	Project short name	CORE
	URL	http://zil.ipipan.waw.pl/CORE
	Funding type	National funds
	Funder	National Science Centre (100%)
	Country	Poland
	Start date	2011-04-18
	End date	2014-04-17
Creation start date	2011-05-01	

Texts

Media type	text
-------------------	------

Linguality type	Monolingual	
Languages	Polish	
	Language ID	pl
Modality	Modality type	Spoken language
		Written language
Size	503985 tokens	
Character encoding	UTF-8	
Annotation	Semantic annotation – named entities	
	Annotation standoff	True
	Segmentation level	Word group
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Automatic
	Annotation mode details	named entities (person names, organizations, locations compatible with NKJP hierarchy) detected by Nerf
	Annotation tool	Nerf, a named entity recognizer for Polish
	Start date	2012-01-01
	Morphosyntactic annotation – POS tagging	
	Annotation standoff	True
	Segmentation level	Word
	Format	text/xml
	Tagset	NKJP tagset
	Tagset language id	Pl
	Tagset language name	Polish
	Conformance to standards best practices	TEI
	Annotation mode	Automatic
	Annotation mode details	MSD and POS tag variants (all available morphosyntactic interpretations) output by Morfeusz, then disambiguated by Pantera

	tagger
Annotation tool	Morfeusz SGJP, a tokenizer, morphological analyzer and lemmatizer for Polish Pantera, a Brill tagger for Polish
Start date	2012-01-01
Segmentation	
Annotation standoff	True
Segmentation level	Sentence Word
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Mixed
Annotation mode details	When required for the purpose of coreference annotation, sentence and word segmentation output by Pantera was corrected manually
Annotation tool	Pantera, a Brill tagger for Polish
Start date	2012-01-01
Lemmatization	
Annotation standoff	False
Segmentation level	Word
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation mode details	lemma variants (all available interpretations) output by Morfeusz, then disambiguated by Pantera tagger
Annotation tool	Morfeusz SGJP, a tokenizer, morphological analyzer and lemmatizer for Polish Pantera, a Brill tagger for Polish
Start date	2012-01-01
Structural annotation	
Annotation standoff	True
Segmentation	Word

level	
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation mode details	syntactic words (word-like compounds) detected by Spejd with NKJP shallow parsing grammar; see NKJP documentation for details
Annotation tool	Spejd, a shallow parser of Polish
Start date	2012-01-01
Syntactic annotation – shallow parsing	
Annotation standoff	True
Segmentation level	Word group
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation mode details	syntactic groups (phrase-like constructs) detected by Spejd with NKJP shallow parsing grammar; see NKJP documentation for details
Annotation tool	Spejd, a shallow parser of Polish
Start date	2012-01-01
Semantic annotation – entity mentions	
Annotation standoff	True
Segmentation level	Word group
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Mixed
Annotation mode details	manual annotation with automatic preannotation
Start date	2012-01-01

	Discourse annotation – coreference	
	Annotation standoff	True
	Segmentation level	Other
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Mixed
	Annotation mode details	manual annotation with automatic preannotation
	Start date	2012-01-01
Creation	Original source	http://nkjp.pl
	Creation mode	Mixed
	Creation tools	Morfeusz SGJP, a tokenizer, morphological analyzer and lemmatizer for Polish Pantera, a Brill tagger for Polish Spejd, a shallow parser of Polish Nerf, a named entity recognizer for Polish

4.39. Polish Coreference Tools

General Information

Description	The Polish Coreference Tools is a suite of tools created during the "Computer-based methods for coreference resolution in Polish texts" project. It is going to contain the tool used for manually annotation of the Polish Coreference Corpus and all the automatic coreference resolution tools created during the project. Currently the suite contains the automatic mention detection and coreference resolution rule-based tool, which was used for preannotation of the PCC.
Identifier	439
Resource type	Tool/service
Tool/service type	Suite of tools
URL	http://zil.ipipan.waw.pl/PolishCoreferenceTools
Version	0.5
Last update	2013-01-08

Contacts

Maciej Ogrodniczuk	

Position	Assistant Professor
Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Mateusz Kopeć	
Contact	m.kopec@ipipan.waw.pl http://zil.ipipan.waw.pl/MateuszKopec
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – unrestricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

CC-BY	
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/PolishCoreferenceTools

Metadata

Creation date	2013-01-08	
Metadata creators	Mateusz Kopeć	
	Contact	m.kopec@ipipan.waw.pl http://zil.ipipan.waw.pl/MateuszKopec
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Metadata last date updated	2013-01-22
-----------------------------------	------------

Resource creation

Resource creator	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu
Funding projects	Computer-based methods for coreference resolution in Polish texts	
	Project short name	CORE
	URL	http://zil.ipipan.waw.pl/CORE
	Funding type	National funds
	Funder	National Science Centre (100%)
	Country	Poland
	Start date	2011-04-18
	End date	2014-04-17
Creation start date	2011-05-01	

Tool/service

Tool/service type	Suite of tools
Language dependent	True

4.40. Syntactic-Generative Dictionary of Polish Verbs

General Information

Short name	Syntactic-Generative Dictionary of Polish Verbs
Description	Syntactic-Generative Dictionary of Polish Verbs (original version created by Kazimierz Polański) is a description of argument structure of 10559 verbs in Polish. It has been called "syntactic-generative" because it does not provide a full description of all Polish verbs, but focuses only on their syntactic behavior while ignoring inflection, word formation and phonology. Semantic information is also reduced to deal only with cases where multiple meanings of given verb imply different sentence structures.
Identifier	440

Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://zil.ipipan.waw.pl/PoliMorf
Version	1.0

Contacts

Maciej Ogrodniczuk	
Position	Assistant Professor
Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Michal Lenart	
Position	Programmer
Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@ipipan.waw.pl http://zil.ipipan.waw.pl/MichalLenart
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – unrestricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

BSD-style

Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/SGDPV	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2011-11-25	
Metadata creators	Michal Lenart	
	Position	Programmer
	Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@ipipan.waw.pl http://zil.ipipan.waw.pl/MichalLenart
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-23	

Resource creation

Resource creator	Michal Lenart	
	Position	Programmer
	Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@ipipan.waw.pl http://zil.ipipan.waw.pl/MichalLenart
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group

	ipi@ipipan.waw.pl	
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
	Creation start date	2012-12-01

Resource documentation

Reports	PDF available at http://zil.ipipan.waw.pl/SlownikPolanskiego
Tool documentation type	Manual Online

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Syntax
	Conformance to standards best practices	TEI_P5

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	latin
Modality	Modality type	Written language
Size	10559 entries	

Character encoding	UTF-8
--------------------	-------

4.41. Manually aligned CES Polish-English parallel corpus

General Information

Short name	CESCorpus
Description	A corpus of the Centre for Eastern Studies (CES) texts. This resource contains 56 Polish-English texts (6 CES reports, 28 issues of CES studies and 22 issues of the CES publication "Point of View") licensed under the CC-BY-NC license. The texts have been aligned manually on the sentence level using the MemoQ software. The resource is provided as TEI P5-compliant XML files with custom extensions and in the XLIFF and TMX formats.
Identifier	441
Resource type	Corpus
URL	http://clip.ipipan.waw.pl/CESCorpus
Version	1.0
Revision	compilation of the corpus
Last update	2012-12-31

Contacts

Adam Przepiórkowski	
Position	Professor, Head of the Linguistic Engineering Group
Contact	Jana Kazimierza 5 01-248 Warsaw adam.przepiorkowski@ipipan.waw.pl http://zil.ipipan.waw.pl/AdamPrzepiorkowski
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – restricted use	
IPR holder	Ośrodek Studiów Wschodnich	
	Contact	ul. Koszykowa 6a 00-564 Warszawa info@osw.waw.pl http://www.osw.waw.pl
Availability start date	2012-06-30	

Licences

CC-BY-NC		
Restrictions of use	Academic – non-commercial use Attribution	
Access medium	Downloadable	
Download location	http://clip.ipipan.waw.pl/CESCorpus	
Attribution text	Pęzik P., Ogrodniczuk M., Przepiórkowski A (2011). Parallel and spoken corpora in an open repository of Polish language resources. Human Language Technologies as a Challenge for Computer Science and Linguistics. LTC Poznań 2011.	
Signatories	Ośrodek Studiów Wschodnich	
	Contact	ul. Koszykowa 6a 00-564 Warszawa info@osw.waw.pl http://www.osw.waw.pl
Distribution rights holder	Ośrodek Studiów Wschodnich	
	Contact	ul. Koszykowa 6a 00-564 Warszawa info@osw.waw.pl http://www.osw.waw.pl

Metadata

Creation date	2012-12-31	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Łukasz Drózd	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	Institute of Computer Science, Polish Academy of Sciences

	Linguistic Engineering Group ipi@ipipan.waw.pl
Metadata language name	English
Metadata language ID	en

Validation

Validated	True	
Type	Formal	
Mode	Automatic	
Details	Validation of the XML structure in accordance with the TEI P5, XLIFF and TMX schemas with custom extensions.	
Extent	Full	
Size	1432000 words	
Validation report	Reports	All files are valid XML conforming to the TEI P5 and XLIFF schemas.
Tool	xmllint	
Validator	Maciej Ogrodniczuk	
	Position	Assistant professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Łukasz Drózdź	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Resource creation

Resource creator	Maciej Ogrodniczuk	
	Position	Assistant professor

	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Start date	2011-02-01
	End date	2013-01-31
Creation start date		2011-08-01
Creation end date		2012-06-30

Resource documentation

Reports	http://clip.ipipan.waw.pl/CESCorpus
----------------	---

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Multilinguality type details	The texts were manually aligned on a sentence level using the MemoQ software (http://kilgray.com/products/memoq).	
Languages	English	
	Language ID	en
	Language script	Latn
	Size	800000 words
	Polish	
	Language ID	pl
	Language script	Latn
	Size	661000 words
Modality	Modality type	Written language
	Size	56 texts
Size	56 texts, 1445000 words	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Segmentation	
	Annotation standoff	False
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI_P5
	Annotation mode	Automatic
	Annotation tool	memoQ (http://kilgray.com/products/memoq)
	Start date	2012-06-01
	End date	2012-11-30
	Size	56 texts
	Annotators	Maciej Ogrodniczuk

		Position	Assistant professor
		Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
		Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
		Łukasz Dróżdż	
		Position	IT specialist
		Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
		Alignment	
		Segmentation level	Sentence
		Format	text/xml
		Conformance to standards best practices	TEI
		Annotation mode	Manual
		Annotation tool	memoQ (http://kilgray.com/products/memoq)
		Start date	2012-06-01
		End date	2012-11-30
		Size	56 texts
		Annotators	Maciej Ogrodniczuk
		Position	Assistant professor
		Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
		Organization	Institute of Computer Science, Polish Academy of Sciences

		Linguistic Engineering Group ipi@ipipan.waw.pl	
		Łukasz Dróżdż	
		Position	IT specialist
		Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Domains	politics		
Time coverages	2003-2012		
Geographic coverages	Europe, Asia		
Creation	Original source	http://www.osw.waw.pl/	
	Creation mode	Manual	
	Creation mode details	The texts were acquired as PDF and converted to plain text. Segmentation and manual alignment were performed using memoQ. Care was taken to represent all non-trivial translation equivalence types.	
	Creation tools	WebLign (http://code.google.com/p/weblign)	

4.42. Polish lexicon for OpenCyc

General Information

Short name	PolishOpenCyc
Description	The Polish OpenCyc lexicon is a set of mappings between OpenCyc symbols and their Polish counterparts. It might be thought of as a English-Polish dictionary, but since Cyc contains many abstract concepts, it is better viewed as a set of Polish names for the concepts that are present in Cyc. Due to the fact, that the latest OpenCyc version contains more than 200 thousands of concepts, the mapping covers only part of it -- the concepts that are also present in Umbel, an ontology that was developed on the basis of Cyc, but is devoid of many application-specific Cyc concepts. Still many concepts that are present in Umbel lack mapping, since they are very specific and doesn't translate into Polish (e.g. names of species found only in North America). As a result the current mapping contains approx. 16 thousands of translations for approx. 14 thousands of concepts.
Identifier	442
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Lexicon
URL	http://zil.ipipan.waw.pl/PolishOpenCyc

Version	0.2
----------------	-----

Contacts

Aleksander Pohl	
Position	Assistant Lecturer
Contact	Lojasiewicza 4 30-348 Krakow apohllo@o2.pl http://apohllo.pl

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

CC-BY	
Restrictions of use	Attribution
Access medium	Downloadable
Download location	https://github.com/apohllo/polish-cyc
Fee	free of charge

Metadata

Creation date	2013-01-18	
Metadata creators	Aleksander Pohl	
	Position	Assistant Lecturer
	Contact	Lojasiewicza 4 30-348 Krakow apohllo@o2.pl http://apohllo.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-18	

Resource creation

Funding projects	Central and South-East European Resources	
	Project short	CESAR

	name	
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-05-01	

Resource documentation

Reports	<p>Aleksander Pohl, The Polish Cyc lexicon as a bridge between Polish language and the Semantic Web, Proceedings of the International Multiconference on Computer Science and Information Technology, IEEE, pp. 485-492 ISBN: 978-83-60810-22-4, ISSN: 1896-7094</p> <p>Aleksander Pohl, The Semi-automatic Construction of the Polish Cyc Lexicon, Investigationes Linguisticae, vol. XXI, pp. 17-38, ISSN 1733-1757</p> <p>Aleksander Pohl, Automatic Construction of the Polish Nominal Lexicon for the OpenCyc Ontology, Recent Advances in Intelligent Information Systems, Mieczysław A. Kłopotek et al red., Oficyna Wydawnicza Exit: Warszawa, pp. 51-64, ISBN: 978-83-60434-59-8</p>
----------------	---

Lexical conceptual resource

Lexical conceptual resource type	Lexicon	
Lexical conceptual resource encoding	Encoding level	Semantics
	Linguistic information	<p>Lemma</p> <p>Lemma – multi word units</p> <p>Lemma – variants</p> <p>Case</p> <p>Degree</p> <p>Gender</p> <p>Number</p> <p>Part of speech</p> <p>Translation equivalent</p>
	External ref	http://sw.opencyc.org
Creation	Original source	http://opencyc.org
	Creation mode	Interactive
	Creation tools	https://github.com/apohllo/cyc-mapping

Texts

Media type	text	
Linguality type	Bilingual	
Languages	Polish	
	Language ID	PL
	Language script	Latn
	English	
	Language ID	EN
	Language script	Latn
Size	2748 kb, 13925 entries	
Character encoding	UTF-8	

4.43. TAG grammar for Polish

General Information

Description	A TAG (Tree Adjoining Grammar) extracted automatically from Składnica constituency treebank. The grammar and lexicon are in XMG (http://wiki.loria.fr/wiki/XMG/Documentation) and LEX2ALL (http://wiki.loria.fr/wiki/LEX2ALL) formats respectively. The grammar contains 2802 elementary tree families (1825 initial trees and 977 auxiliary trees). The lexicon contains 11515 lexemes, anchoring a total of 23399 trees (one lexeme can serve as a lexical anchor to more than one tree, e.g. in case of verbs with more than one possible valence frame).
Identifier	443
Resource type	Language description
Language description type	Grammar
URL	http://zil.ipipan.waw.pl/pITAG

Contacts

Katarzyna Krasnowska	
Contact	kasia.krasnowska@gmail.com

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

GPL	
Restrictions of use	Attribution
Access medium	Downloadable

Metadata

Creation date	2013-01-17	
Metadata creators	Katarzyna Krasnowska	
	Contact	kasia.krasnowska@gmail.com
Metadata language name	English	
Metadata language ID	en	
Metadata last date updated	2013-01-17	

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-05-01	

Language description

Language description type	Grammar	
Language description encoding	Encoding level	Syntax
	Theoretic model	Tree Adjoining Grammar
Texts	Media type	text

	Linguality type	Monolingual	
	Languages	Polish	
		Language ID	pl

Texts

Media type	text		
Linguality type	Monolingual		
Languages	Polish		
	Language ID	pl	

4.44. Anotatornia

General Information

Short name	Anotatornia		
Description	Anotatornia is a tool for the manual on-line annotation of corpora at various linguistic levels. The levels currently implemented are: word-level and sentence-level segmentation, morphosyntax, word sense disambiguation. Anotatornia implements sophisticated mechanisms of the management of texts, annotators and conflicts.		
Identifier	444		
Resource type	Tool/service		
Tool/service type	Nlp development environment		
URL	http://zil.ipipan.waw.pl/Anotatornia		
Version	0.1		
Last update	2013-01-17		

Contacts

Michał Lenart	
Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@ipipan.waw.pl

Distribution

Availability	Available – restricted use		
IPR holder	Institute of Computer Science, Polish Academy of Sciences		
	Short name	IPI PAN	
	Department	Linguistic Engineering Group	

	name	
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Licences

GPL		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/Anotatoria	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2013-01-16	
Metadata creators	Michal Lenart	
	Position	Assistant
	Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@ipipan.waw.pl
Source	NKJP	
Metadata language ID	en	
Metadata last date updated	2013-01-17	

Usage

Foreseen use	NLP applications
NLP-specific use	Annotation

Actual uses	NLP applications	
	NLP-specific use	Annotation
	Reports	Adam Przepiórkowski and Grzegorz Murzynowski. Manual annotation of the National Corpus of Polish with Anotatornia. In Stanisław Goźdz-Roszkowski, editor, Explorations across Languages and Corpora: PALC 2009, pages 95–103, Frankfurt am Main, 2011. Peter Lang.
	Usage project	National Corpus of Polish
		Project short name NKJP
		URL http://nkjp.pl/
		Funding type National funds
		Funder Polish Ministry of Science and Higher Education
		Country Poland
		Start date 2007-01-01
		End date 2011-06-30

Resource creation

Resource creator	Grzegorz Murzynowski	
	Contact	natror@o2.pl
Funding projects	National Corpus of Polish	
	Project short name	NKJP
	URL	http://nkjp.pl/
	Funding type	National funds
	Funder	Polish Ministry of Science and Higher Education
	Country	Poland
	Start date	2007-01-01
	End date	2011-06-30
Creation start date	2010-03-01	

Resource documentation

Documents	Manual	
	Title	Anotatornia - readme
Tool documentation type	Help functions	

Tool/service

Tool/service type	Nlp development environment	
Language dependent	True	
Input	Media type	text
	Modality type	Written language
Output	Lemmatization	
	Media type	text
	Modality type	Written language
Operating system	Linux	
Required software	Ruby v \geq 1.8 SQLite3 v \geq 3.3.6 sqlite3-ruby v \geq 1.2.4 mongrel v \geq 1.1.5 Rails v1.2.4	
Tool/service evaluation	Evaluated	False

4.45. Constrained Conditional Random Fields Tagging Tool

General Information

Short name	Concraft
Description	Concraft is a statistical tool for morphosyntactic disambiguation developed as a part of the CESAR project. It is based on conditional random fields (CRFs) extended with additional, position-wise restrictions on the output domain, which are used to impose consistency between the modeled label sequences and morphosyntactic analysis results both at the level of decoding and, more importantly, in parameters estimation process. The problem of morphosyntactic disambiguation is decomposed into two consecutive stages of the context-sensitive morphosyntactic guessing and the disambiguation proper. The tool is currently adapted to the Polish language and resources, but the method and the library should be applicable to at least other highly inflected languages.
Identifier	445
Resource type	Tool/service
Tool/service type	Tool
URL	http://zil.ipipan.waw.pl/Concraft
Version	0.1
Last update	2013-01-15

Contacts

Jakub Waszczuk	
Contact	Jana Kazimierza 5

	01-248 Warsaw waszczuk.kuba@gmail.com
--	---

Distribution

Availability	Available – restricted use
--------------	----------------------------

Licences

BSD	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/Concraft
Fee	free of charge

Metadata

Creation date	2013-01-15	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Jakub Waszczuk	
	Contact	Jana Kazimierza 5 01-248 Warsaw waszczuk.kuba@gmail.com
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-15	

Usage

Foreseen use	NLP applications	
NLP-specific use	Morphosyntactic tagging	
Actual uses	NLP applications	
	NLP-specific use	Morphosyntactic tagging

	Reports	Jakub Waszczuk. 2012. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In Proceedings of COLING 2012, Mumbai, India.	
	Usage project	Central and South-East European Resources	
		Project short name	CESAR
		URL	http://www.cesar-project.net
		Funding type	EU funds National funds Own funds
		Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
		Country	Poland
		Start date	2011-02-01
		End date	2013-01-31
	Actual use details	Concraft evaluation and comparison to other state-of-the-art taggers for Polish.	

Resource creation

Resource creator	Jakub Waszczuk	
	Contact	Jana Kazimierza 5 01-248 Warsaw waszczuk.kuba@gmail.com
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-02-01	

Resource documentation

Reports	Jakub Waszczuk. 2012. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In Proceedings of COLING 2012, Mumbai, India.
Tool documentation type	Online

Tool/service

Tool/service type	Tool	
Language dependent	False	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	OS-independent	
Required software	The Glasgow Haskell Compiler (version 7.0.4 or higher)	
Tool/service evaluation	Evaluated	True
	Level	Diagnostic
	Type	Black box
	Criteria	Intrinsic
	Measure	Automatic
	Reports	10-fold cross-validation on the one-million-word, balanced subcorpus of the National Corpus of Polish. Results and more detailed description can be found in: Jakub Waszczuk. 2012. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In Proceedings of COLING 2012, Mumbai, India.
	Details	Evaluation has been performed with respect to the guidelines described in: Radziszewski, A. and Acedański, S. (2012). Taggers gonna tag: an argument against evaluating disambiguation capacities of morphosyntactic taggers. In Proceedings of TSD 2012, LNCS. Springer-Verlag.
	Evaluators	Jakub Waszczuk
		Contact Jana Kazimierza 5 01-248 Warsaw waszczuk.kuba@gmail.com
Tool/service creation	Implementation language	Haskell

	Formalism	Constrained Conditional Random Fields Tiered Tagging
--	------------------	---

4.46. Multiservice

General Information

Short name	Multiservice
Description	Multiservice is a web service integration platform for Polish linguistics resources. It is designed for chaining execution of linguistic tools. Processing is triggered by request sent to the web service. Requests are enqueued and handled in asynchronous manner. It is accessible through portable and simple SOAP-based API. New resources can be plugged-in relatively easily using unified API based on Apache Thrift framework. Online demo providing graphical representation of request results is also included.
Identifier	446
Resource type	Tool/service
Tool/service type	Service
URL	http://zil.ipipan.waw.pl/Multiservice
Version	0.2
Last update	2013-01-22

Contacts

Michał Lenart	
Position	Programmer
Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@gmail.com
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – restricted use	
IPR holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl

	http://www.ipipan.eu
--	---

Licences

GPL		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://zil.ipipan.waw.pl/Multiservice	
Fee	free of charge	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Short name	IPI PAN
	Department name	Linguistic Engineering Group
	Contact	Jana Kazimierza 5 01-248 Warsaw ipi@ipipan.waw.pl http://www.ipipan.eu

Metadata

Creation date	2013-01-22	
Metadata creators	Maciej Ogrodniczuk	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Michal Lenart	
	Position	Programmer
	Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@gmail.com
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Source	CESAR	
Metadata language	en	

ID	
Metadata last date updated	2013-01-22

Usage

Foreseen use	NLP applications	
NLP-specific use	Annotation	
Actual uses	NLP applications	
	NLP-specific use	Annotation

Resource creation

Resource creator	Michał Lenart	
	Position	Programmer
	Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@gmail.com http://zil.ipipan.waw.pl/MichalLenart
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2011-06-01	

Resource documentation

Reports	Maciej Ogrodniczuk and Michał Lenart. Web Service integration platform for Polish linguistic resources. In Proceedings of the Eighth International Conference on Language
----------------	---

	Resources and Evaluation, LREC 2012, pages 1164–1168, Istanbul, Turkey, 2012. ELRA.
Tool documentation type	Help functions

Tool/service

Tool/service type	Service	
Tool/service subtype	Web service accessible through SOAP XML messages.	
Language dependent	False	
Input	UTF-8	
	Media type	text
	Modality type	Written language
Output	UTF-8	
	Media type	text
	Modality type	Written language
Operating system	OS-independent	
Tool/service evaluation	Evaluated	False
Tool/service creation	Implementation language	Java Python C++ Javascript

4.47. DBpedia resource classification into the OpenCyc taxonomy

General Information

Short name	DBpediaCycTypes
Description	The files contain the classification of the DBpedia resources into the taxonomy of OpenCyc. They were obtained from the Wikipedia infoboxes, introductory sentences, categories and direct mapping between Wikipedia and Cyc. The results were cross-validated and conflicts were resolved using the Cyc built-in inconsistency detection mechanism. Only the most specific types are provided. All more general types might be retrived using Cyc's built-in taxonomy traversal methods (all-genls function in particular).
Identifier	447
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Ontology
URL	http://klon.wzks.uj.edu.pl/wiki-types/
Version	0.1

Contacts

Pohl Aleksander	
Position	Assistant Lecturer
Contact	Lojasiewicza 4 30-348 Krakow apohllo@o2.pl http://apohllo.pl

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

CC-BY-SA	
Restrictions of use	Attribution Share alike
Access medium	Downloadable
Download location	http://klon.wzks.uj.edu.pl/wiki-types/
Fee	free of charge

Metadata

Creation date	2013-01-27	
Metadata creators	Pohl Aleksander	
	Position	Assistant Lecturer
	Contact	Lojasiewicza 4 30-348 Krakow apohllo@o2.pl http://apohllo.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-27	

Resource documentation

Reports	Aleksander Pohl, Classifying the Wikipedia Articles into the OpenCyc Taxonomy, Proceedings of the Web of Linked Entities Workshop in conjunction with the 11th International Semantic Web Conference, Giuseppe Rizzo, Pablo Mendes, Eric Charton,
----------------	---

Lexical conceptual resource

Lexical conceptual resource type	Ontology	
Lexical conceptual resource encoding	Encoding level	Semantics
	Linguistic information	Semantics – relations – hyperonyms
	External ref	http://sw.opencyc.org
Creation	Original source	http://opencyc.org http://en.wikipedia.org http://dbpedia.org
	Creation mode	Automatic

Texts

Media type	text	
Linguality type	Monolingual	
Languages	English	
	Language ID	EN
	Language script	Latn
Size	8.19 mb, 2221000 entries	
Character encoding	UTF-8	

4.48. DBPedia Extender

General Information

Short name	DBPediaExtender
Description	The DBPediaExtender is an information extraction system that extends an existing ontology of geographical entities by extracting information from text. The system uses distant supervision learning – training data is constructed based on matches between values from a knowledge base (DBPedia) and Wikipedia articles. The system was run on the Polish versions of DBPedia and Wikipedia and extracted more than 44 thousand RDF triples expressing relations between geographic entities from Polish Wikipedia.
Identifier	448
Resource type	Tool/service
Tool/service type	Tool
URL	http://zil.ipipan.waw.pl/DBPediaExtender
Version	0.1

Last update	2013-01-23
--------------------	------------

Contacts

Marcin Zajac	
Contact	Jana Kazimierza 5 01-248 Warsaw mrcnzajac@gmail.com

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

GPL	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/DBPediaExtender
Fee	free of charge

Metadata

Creation date	2013-01-23	
Metadata creators	Marcin Zajac	
	Contact	Jana Kazimierza 5 01-248 Warsaw mrcnzajac@gmail.com
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-23	

Usage

Foreseen use	NLP applications	
NLP-specific use	Information extraction	
Actual uses	NLP applications	
	Actual use details	Extraction of semantic relations from Polish Wikipedia.

Resource creation

Resource creator	Marcin Zając	
	Contact	Jana Kazimierza 5 01-248 Warsaw mrcnzajac@gmail.com
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-02-01	

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Text/plain	
	Media type	text
	Modality type	Written language
Output	Semantic annotation	
	Media type	text
	Modality type	Written language
Operating system	Linux	
Required software	Python >= 2.6 scikit-learn crfsuite pantera-tagger OpenLink Virtuoso (Open-Source Edition)	
Tool/service	Implementation	Python

creation	language	
	Formalism	Distant supervision learning Conditional random fields Support vector machines

4.49. LFG Treebank for Polish

General Information

Short name	pol-lfg-treebank
Description	A collection of sentences from the Polish National Corpus, parsed with the LFG grammar, represented as syntactic trees and analysed for grammatical functions.
Identifier	449
Resource type	Language description
Language description type	Other
URL	http://zil.ipipan.waw.pl/LFG
Version	1.0
Last update	2013-01-23

Contacts

Anna Kibort	
Position	structure bank author
Contact	Jana Kazimierza 5 01-248 Warsaw ak243@cam.ac.uk

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

GPL	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/LFG
Fee	free of charge

Metadata

--	--

Creation date	2013-01-23	
Metadata creators	Anna Kibort	
	Contact	Jana Kazimierza 5 01-248 Warsaw ak243@cam.ac.uk
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-23	

Usage

Foreseen use	NLP applications
NLP-specific use	Parsing

Resource creation

Resource creator	Anna Kibort	
	Contact	Jana Kazimierza 5 01-248 Warsaw ak243@cam.ac.uk
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-11-09	

Language description

Language description type	Other
----------------------------------	-------

Language description encoding	Encoding level	Syntax	
	Theoretic model	LFG (Lexical Functional Grammar) grammars minimally provide two levels of representation: constituent structure (c-structure) produced by context-free phrase structure rules and functional structure (f-structure) created by functional descriptions.	
Texts	Media type	text	
	Linguality type	Monolingual	
	Languages	Polish	
		Language ID	PL
		Language script	UTF-8
	Modality	Modality type	Written language
	Size	837 sentences	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	UTF-8
Modality	Modality type	Written language
Size	837 sentences	

4.50. NKJP1mEcono corpus

General Information

Short name	NKJP1mEcono
Description	Economy-related subcorpus of the National Corpus of Polish, containing manually created sense annotation layer.
Identifier	450
Resource type	Corpus
URL	http://zil.ipipan.waw.pl/NKJP1mEcono
Version	1.0

Contacts

--

Lukasz Kobyliński	
Position	Research Assistant
Contact	Jana Kazimierza 5 01-248 Warsaw lkobylnski@ipipan.waw.pl http://zil.ipipan.waw.pl/LukaszKobylnski

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

GPL	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/NKJP1mEcono?action=AttachFile&do=get&target=nkjp1m-econo.zip
Fee	free of charge

Metadata

Creation date	2013-01-23	
Metadata creators	Lukasz Kobyliński	
	Position	Research Assistant
	Contact	Jana Kazimierza 5 01-248 Warsaw lkobylnski@ipipan.waw.pl http://zil.ipipan.waw.pl/LukaszKobylnski
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-23	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL

Size	11 mb, 87816 words	
Annotation	Segmentation	
	Annotation standoff	True
	Segmentation level	Paragraph
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode details	inherited from source corpus (no need to generate new segmentation when sampling)
	Start date	2010-02-01
	End date	2010-02-28
	Segmentation	
	Annotation standoff	True
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Automatic
	Annotation tool	TaKIPI 1.8
	Start date	2010-02-01
	End date	2010-02-28
	Segmentation	
	Annotation standoff	True
	Segmentation level	Word
	Format	text/xml
	Tagset	NKJP tagset
	Conformance to standards best practices	TEI
	Annotation mode	Automatic

Annotation tool	TaKIPI 1.8
Start date	2010-02-01
End date	2010-02-28
Lemmatization	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation tool	TaKIPI 1.8
Start date	2010-02-01
End date	2010-02-28
Morphosyntactic annotation - below POS tagging	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation tool	TaKIPI 1.8
Start date	2010-02-01
End date	2010-02-28
Morphosyntactic annotation – POS tagging	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to	TEI

	standards best practices	
	Annotation mode	Automatic
	Annotation tool	TaKIPI 1.8
	Start date	2010-02-01
	End date	2010-02-28
	Semantic annotation – word senses	
	Annotation standoff	True
	Segmentation level	Word
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Manual
	Annotation mode details	manually disambiguated using AnotEk
	Annotation tool	AnotEk
	Start date	2010-06-01
	End date	2010-09-01
Creation	Original source	1 million subcorpus of National Corpus of Polish (1MNKJP)
	Creation mode	Mixed
	Creation mode details	The corpus has been created by selecting economy-related paragraphs from the 1M subcorpus of the National Corpus of Polish. Manual word sense annotation has been created by linguists using AnotEk.
	Creation tools	TaKIPI 1.8 Java code AnotEk 1.0

4.51. gpwEcono corpus

General Information

Short name	gpwEcono
Description	A corpus of Polish language stock market reports, with manual annotation on the word sense layer and automatic morphosyntactic annotation, TEI format.
Identifier	451

Resource type	Corpus
URL	http://zil.ipipan.waw.pl/gpwEcono
Version	1.0

Contacts

Lukasz Kobyliński	
Position	Research Assistant
Contact	Jana Kazimierza 5 01-248 Warsaw lkobylnski@ipipan.waw.pl http://zil.ipipan.waw.pl/LukaszKobylnski

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

GPL	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/gpwEcono?action=AttachFile&do=get&target=gpw-econo.zip
Fee	free of charge

Metadata

Creation date	2013-01-23	
Metadata creators	Lukasz Kobyliński	
	Position	Research Assistant
	Contact	Jana Kazimierza 5 01-248 Warsaw lkobylnski@ipipan.waw.pl http://zil.ipipan.waw.pl/LukaszKobylnski
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-23	

Texts

--	--

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
Size	20 mb, 282366 words	
Annotation	Segmentation	
	Annotation standoff	True
	Segmentation level	Paragraph
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode details	inherited from source corpus (no need to generate new segmentation when sampling)
	Start date	2010-02-01
	End date	2010-02-28
	Segmentation	
	Annotation standoff	True
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Automatic
	Annotation tool	TaKIPI 1.8
	Start date	2010-02-01
	End date	2010-02-28
	Segmentation	
	Annotation standoff	True
	Segmentation level	Word
	Format	text/xml
	Tagset	NKJP tagset
	Conformance to	TEI

standards best practices	
Annotation mode	Automatic
Annotation tool	TaKIPI 1.8
Start date	2010-02-01
End date	2010-02-28
Lemmatization	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation tool	TaKIPI 1.8
Start date	2010-02-01
End date	2010-02-28
Morphosyntactic annotation - below POS tagging	
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Tagset	NKJP tagset
Conformance to standards best practices	TEI
Annotation mode	Automatic
Annotation tool	TaKIPI 1.8
Start date	2010-02-01
End date	2010-02-28
Morphosyntactic annotation – POS tagging	
Annotation standoff	True
Segmentation	Word

	level	
	Format	text/xml
	Tagset	NKJP tagset
	Conformance to standards best practices	TEI
	Annotation mode	Automatic
	Annotation tool	TaKIPI 1.8
	Start date	2010-02-01
	End date	2010-02-28
	Semantic annotation – word senses	
	Annotation standoff	True
	Segmentation level	Word
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Manual
	Annotation mode details	manually disambiguated using AnotEk
	Annotation tool	AnotEk
	Start date	2010-06-01
	End date	2010-09-01
Creation	Original source	http://www.gpwinfostrefa.pl/
	Creation mode	Mixed
	Creation mode details	The corpus has been created from stock market reports. Morphosyntactic annotation has been done using the TaKIPI tagger. AnotEk was used for manual word sense annotation.
	Creation tools	TaKIPI 1.8 Java code AnotEk 1.0

4.52. Summary Annotation Tools

General Information

--	--

Description	A set of 4 similar desktop applications for summary annotation. These 4 applications facilitate manual annotation of clauses, extract (unconstrained) summaries, extract clause-grained summaries and abstract summaries.
Identifier	452
Resource type	Tool/service
Tool/service type	Nlp development environment
URL	http://zil.ipipan.waw.pl/SummaryAnnotationTools

Contacts

Mateusz Kopec	
Contact	m.kopec@ipipan.waw.pl http://zil.ipipan.waw.pl/MateuszKopec
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

CC-BY	
Download location	http://zil.ipipan.waw.pl/SummaryAnnotationTools

Metadata

Creation date	2013-01-23	
Metadata creators	Mateusz Kopec	
	Contact	m.kopec@ipipan.waw.pl http://zil.ipipan.waw.pl/MateuszKopec
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Metadata last date updated	2013-01-23	

Tool/service

Tool/service type	Nlp development environment
Language dependent	False

4.53. DistSys

General Information

Description	A distribution system for texts for any kind of manual annotation. Facilitates random assignment of texts to the annotators, distribution of these texts and finally allows to gather them to form a manually annotated corpus.
Identifier	453
Resource type	Tool/service
Tool/service type	Nlp development environment
URL	http://zil.ipipan.waw.pl/DistSys

Contacts

Mateusz Kopec	
Contact	m.kopec@ipipan.waw.pl http://zil.ipipan.waw.pl/MateuszKopec
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Jakub Waszczuk	
Contact	Jana Kazimierza 5 01-248 Warsaw waszczuk.kuba@gmail.com
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Under negotiation
---------------------	-------------------

Metadata

Creation date	2013-01-23	
Metadata creators	Mateusz Kopec	
	Contact	m.kopec@ipipan.waw.pl http://zil.ipipan.waw.pl/MateuszKopec
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Metadata last date updated	2013-01-23	

Tool/service

Tool/service type	Nlp development environment
Language dependent	False

4.54. The Polish SRL corpus

General Information

Description	A part of PELCRA corpus annotated manually with FrameNet semantic roles.
Identifier	454
Resource type	Corpus
URL	http://zil.ipipan.waw.pl/SRL/
Version	1.0

Contacts

Konrad Goluchowski	
Contact	k.goluchowski@phd.ipipan.waw.pl
Organization	Polish Academy of Sciences Institute of Computer Science k.goluchowski@phd.ipipan.waw.pl

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

CC-BY		
Restrictions of use	Academic – non-commercial use Attribution Share alike	
Access medium	Downloadable	
Execution location	http://zil.ipipan.waw.pl/SRL/	
Distribution rights holder	Institute of Computer Science, Polish Academy of Sciences	
	Contact	Jana Kazimierza 5 01-248 Warsaw k.goluchowski@phd.ipipan.waw.pl http://zil.ipipan.waw.pl/SRL/

Metadata

Creation date	2013-01-26	
Metadata creators	Konrad Goluchowski	
	Contact	k.goluchowski@phd.ipipan.waw.pl
	Organization	Polish Academy of Sciences Institute of Computer Science k.goluchowski@phd.ipipan.waw.pl

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
Size	2.4 mb, 266 sentences	
Annotation	Morphosyntactic annotation – POS tagging	
	Segmentation level	Word
	Semantic annotation – semantic roles	
	Segmentation level	Word

4.55. Składnica — a treebank of Polish

General Information

Short name	Składnica
Description	A constituency and dependency treebank of Polish based on sentences from the Polish National Corpus (NKJP), parsed with the Świgr parser, manually verified.
Identifier	455
Resource type	Language description
Language description type	Other
URL	http://zil.ipipan.waw.pl/Skladnica
Version	0.5
Last update	2011-12-05

Contacts

--

Marcin Woliński	
Position	project manager
Contact	Jana Kazimierza 5 01-248 Warsaw wolinski@ipipan.waw.pl

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

GPL	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/Skladnica
Fee	free of charge

Metadata

Creation date	2013-01-23	
Metadata creators	Marcin Woliński	
	Contact	Jana Kazimierza 5 01-248 Warsaw wolinski@ipipan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-23	

Usage

Foreseen use	NLP applications
NLP-specific use	Parsing

Resource creation

Resource creator	Marcin Woliński	
	Contact	Jana Kazimierza 5 01-248 Warsaw

		wolinski@ipipan.waw.pl
Funding projects	Construction of a treebank for Polish using automatic syntactic analysis	
	URL	http://clip.ipipan.waw.pl/Construction of a treebank for Polish using automatic syntactic analysis
	Funding type	National funds
	Funder	Polish Ministry of Science and Higher Education
	Country	Poland
	Start date	2008-10-14
	End date	2011-10-13
Creation start date	2008-11-01	

Language description

Language description type	Other		
Language description encoding	Encoding level	Syntax	
	Theoretic model	DCG grammar	
Texts	Media type	text	
	Linguality type	Monolingual	
	Languages	Polish	
		Language ID	PL
		Language script	UTF-8
	Modality	Modality type	Written language
	Size	8227 sentences	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	UTF-8
Modality	Modality type	Written language
Size	8227 sentences	

4.56. Świgr — a DCG parser of Polish

General Information

Short name	Świgr
Description	Świgr is a parser of Polish using a DCG grammar derived from Marek Świdziński's grammar GFJP (Gramatyka formalna języka polskiego, 1992). Current version includes extensions developed for the Składnica treebank.
Identifier	456
Resource type	Language description
Language description type	Grammar
URL	http://zil.ipipan.waw.pl/ Świgr
Version	1.5
Last update	2011-06-27

Contacts

Marcin Woliński	
Position	author
Contact	Jana Kazimierza 5 01-248 Warsaw wolinski@ipipan.waw.pl

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

GPL	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/ Świgr
Fee	free of charge

Metadata

Creation date	2013-01-23	
Metadata creators	Marcin Woliński	
	Contact	Jana Kazimierza 5 01-248 Warsaw

	wolinski@ipipan.waw.pl
Source	CESAR
Metadata language ID	en
Metadata last date updated	2013-01-23

Usage

Foreseen use	NLP applications
NLP-specific use	Parsing

Resource creation

Resource creator	Marcin Woliński	
	Contact	Jana Kazimierza 5 01-248 Warsaw wolinski@ipipan.waw.pl
Funding projects	Construction of a treebank for Polish using automatic syntactic analysis	
	URL	http://clip.ipipan.waw.pl/Construction of a treebank for Polish using automatic syntactic analysis
	Funding type	National funds
	Funder	Polish Ministry of Science and Higher Education
	Country	Poland
	Start date	2008-10-14
	End date	2011-10-13
Creation start date	2008-11-01	

Language description

Language description type	Grammar		
Language description encoding	Encoding level	Syntax	
	Theoretic model	DCG grammar	
Texts	Media type	text	
	Linguality type	Monolingual	
	Languages	Polish	
		Language ID	PL

		Language script	UTF-8
	Modality	Modality type	Written language
	Size	156197 bytes	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
	Language script	UTF-8
Modality	Modality type	Written language
Size	156197 bytes	

4.57. NKJP Model for TnT Tagger for Polish

General Information

Short name	TnT Tagging Model for Polish
Description	TnT Tagger model used to tag tokenized Polish text. It was trained on the manually annotated Polish National Corpus.
Identifier	457
Resource type	Tool/service
Tool/service type	Tool
URL	http://zil.ipipan.waw.pl/NKJP model for TnT Tagger
Version	0.1
Last update	2013-01-16

Contacts

Marcin Milkowski	
Contact	Jana Kazimierza 5 01-248 Warsaw mmilkows@ifispan.waw.pl

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

BSD	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://zil.ipipan.waw.pl/NKJP model for TnT Tagger
Fee	free of charge

Metadata

Creation date	2013-01-16	
Metadata creators	Marcin Milkowski	
	Position	Assistant
	Contact	Jana Kazimierza 5 01-248 Warsaw mmilkows@ifispan.waw.pl http://zil.ipipan.waw.pl/MarcinMilkowski
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-16	

Usage

Foreseen use	NLP applications	
NLP-specific use	Morphosyntactic tagging	
Actual uses	NLP applications	
	NLP-specific use	Morphosyntactic tagging

Resource creation

Resource creator	Marcin Milkowski	
	Position	Computational linguist
	Contact	Jana Kazimierza 5 01-248 Warsaw mmilkows@ifispan.waw.pl http://zil.ipipan.waw.pl/MarcinMilkowski
Funding projects	Central and South-East European Resources	
	Project short	CESAR

	name	
	URL	http://www.cesar-project.net
	Funding type	EU funds National funds Own funds
	Funder	European Commission (50%) Polish Ministry of Science and Higher Education (40%) Institute of Computer Science, Polish Academy of Sciences (10%)
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2013-01-16	

Resource documentation

Tool documentation type	Help functions
--------------------------------	----------------

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Media type	text
	Modality type	Written language
Output	Media type	text
	Modality type	Written language
Operating system	Linux Mac OS Windows	
Required software	TnT Tagger	
Tool/service evaluation	Evaluated	True
	Level	Diagnostic
	Type	Black box
	Criteria	Intrinsic
	Measure	Automatic
	Reports	The results were evaluated using tnt-diff.
	Details	Polish NKJP model for TnT was trained and tested using 10-fold cross-validation on a million-token manually annotated subcorpus of the

		National Corpus of Polish.	
	Evaluators	Marcin Milkowski	
		Contact	Jana Kazimierza 5 01-248 Warsaw mmilkows@ifspan.waw.pl

4.58. Polish Automatic Collocations Dictionary

General Information

Description	The Polish Automatic Collocations Dictionary has been created by Lexical Computing Ltd. and have been made available to the research community as part of the CESAR initiative.
Identifier	458
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://clip.ipipan.waw.pl/PolishACD

Contacts

Maciej Ogrodniczuk	
Position	Assistant Professor
Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Distribution

Availability	Available – unrestricted use	
IPR holder	Lexical Computing Ltd.	
	Contact	71, Freshfield Road BN2 0BL Brighton inquiries@sketchengine.co.uk http://www.sketchengine.co.uk/
Availability end date	2012-10-26	

Licences

--	--

Access medium	Downloadable
----------------------	--------------

Metadata

Creation date	2013-01-31	
Metadata creators	Michał Lenart	
	Position	Programmer
	Contact	Jana Kazimierza 5 01-248 Warsaw michal.lenart@ipipan.waw.pl http://zil.ipipan.waw.pl/MichalLenart
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2013-01-31	

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary
---	-----------------------------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	pl
Size	30000 entries	

4.59. Polish Corpus of Wrocław University of Technology

General Information

Short name	KPW _r
Description	Polish Corpus of Wrocław University of Technology (PL: Korpus Języka Polskiego Politechniki Wrocławskiej, KPW _r) is a corpus of written and spoken documents available on the Creative Commons license. It comprises of 350k tokens. The texts are divided into 14 categories (blogs, science, stenographic recordings, dialogue, contemporary and old prose, law, long and short press articles, popular science and textbooks, Wikipedia, religion, official and technical texts). The documents are annotated on the level of chunks

	and selected predicate-argument relations, named entities, relations between named entities, anaphora relations and word senses.
Identifier	459
Resource type	Corpus
URL	http://nlp.pwr.wroc.pl/pl/narzedzia-i-zasoby/kpwr

Contacts

Maciej Piasecki	
Contact	maciej.piasecki@pwr.wroc.pl http://www.iis.pwr.wroc.pl/~piasecki
Organization	Institute of Informatics, Wrocław University of Technology Linguistic Engineering Group G 4.19 maciej.piasecki@pwr.wroc.pl

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

CC-BY	
Download location	http://nlp.pwr.wroc.pl/pl/narzedzia-i-zasoby/kpwr

Metadata

Creation date	2013-01-31	
Metadata creators	Marek Miszewski	
	Contact	mawroc@gmail.com
	Organization	Institute of Informatics, Wrocław University of Technology Linguistic Engineering Group G 4.19 maciej.piasecki@pwr.wroc.pl
	Marek Maziarz	
	Contact	marek.maziarz@pwr.wroc.pl http://www.nlp.pwr.wroc.pl/pl/marek-maziarz/4/0/1/show/0/member
	Organization	Institute of Informatics, Wrocław University of Technology Linguistic Engineering Group G 4.19 maciej.piasecki@pwr.wroc.pl
Metadata last date updated	2013-01-31	

Texts

Media type	text
-------------------	------

Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
Size	350k tokens	

5. ULodz resources

5.1. PELCRA Polish-English parallel corpora (CC-BY)

General Information

Short name	PELCRA-PAR-1
Description	A subset of the PELCRA Polish parallel corpora licensed under the CC-BY license. This resource contains 10268 texts from the CORDIS website, 23319 texts from the JRC-Acquis and 4740 texts from the RAPID site. Individual headers may override the licensing information.
Identifier	501
Resource type	Corpus
URL	http://pelcra.pl/res/parallel/pelcra-par-1
Version	1.0
Revision	compilation of the corpus
Last update	2011-09-30

Contacts

Piotr Pęzik	
Position	assistant professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Łukasz Drózdź	
Position	IT specialist
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl

Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
---------------------	---

Distribution

Availability	Available – restricted use	
IPR holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Chair of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Availability start date	2011-11-30	

Licences

CC-BY		
Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://pelcra.pl/res/paralle/pelcra-par-1	
Signatories	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Distribution rights holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Chair of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl

Metadata

Creation date	2011-10-24	
Metadata creators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
	Drózd Łukasz	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Source	CESAR	
Metadata language name	English	
Metadata language ID	en	
Metadata last date updated	2013-01-22	

Validation

Validated	True
Type	Formal
Mode	Automatic
Details	Validation of the XML structure in accordance with the TEI P5 and XLIFF schemas with custom PELCRA extensions.
Extent	Full
Size	75700000 tokens
Tool	xmllint
Validator	Piotr Pęzik

	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl

Resource creation

Resource creator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
	University of Łódź	
	Short name	ULodz
	Department	PELCRA group, Department of English Language and Applied

	name	Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2011-08-01	
Creation end date	2011-09-30	

Resource documentation

Reports	http://pelcra.pl/projects/documentation/
----------------	---

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Multilinguality type details	The texts were aligned on a sentence level using the statistical aligner Maligna (http://align.sourceforge.net).	
Languages	English	
	Language ID	en
	Size	3500000 tokens
	Polish	
	Language ID	pl
	Size	3200000 tokens
Modality	Modality type	Other
Size	6700000 tokens	
Text format	text/xml	

Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Mixed
	Annotation tool	in-house software
	Start date	2011-08-01
	End date	2011-09-30
	Size	6700000 tokens
	Annotators	Piotr Pęzik
		Position assistant professor
		Contact Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
		Drózd Łukasz
		Position IT specialist
		Contact Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
		Alignment
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI

	Annotation mode	Automatic		
	Annotation tool	Maligna		
	Start date	2011-08-01		
	End date	2011-09-30		
	Size	6700000 tokens		
	Annotators	Piotr Pęzik		
		Position	assistant professor	
		Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl	
		Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl	
		Drózd Łukasz		
Position		IT specialist		
Contact		Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl		
Organization		University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl		
Domains	science			
Time coverages	2003-2011			
Geographic coverages	European Union			
Creation	Original source	http://cordis.europa.eu/news/		
	Creation mode	Mixed		
	Creation mode details	Semi-automatic acquisition and processing.		
	Creation tools	in-house software		
Media type	text			
Linguality type	Bilingual			
Multilinguality type	Parallel			
Multilinguality type details	The texts were aligned by the JRC (http://optima.jrc.it) on a sentence level using the statistical aligner hunAlign (http://mokk.bme.hu/resources/hunalign/).			

Languages	English	
	Language ID	en
	Size	32400000 tokens
	Polish	
	Language ID	pl
	Size	28600000 tokens
Modality	Modality type	Other
Size	61000000 tokens	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Mixed
	Annotation tool	unknown
	Start date	2008-05-29
	End date	2008-05-29
	Size	61000000 tokens
	Annotators	Ralf Steinberger
		Position Language Technology Project Manager
		Contact Via E. Fermi 2749 21027 Ispra ralf.steinberger@jrc.ec.europa.eu http://langtech.jrc.ec.europa.eu
		Organization European Commission Joint Research Centre ralf.steinberger@jrc.ec.europa.eu
	Alignment	
	Segmentation level	Sentence
	Format	text/xml
	Conformance to	TEI

	standards best practices	
	Annotation mode	Automatic
	Annotation tool	hunAlign
	Start date	2008-05-29
	End date	2008-05-29
	Size	61000000 tokens
	Annotators	Ralf Steinberger
		Position Language Technology Project Manager
		Contact Via E. Fermi 2749 21027 Ispra ralf.steinberger@jrc.ec.europa.eu http://langtech.jrc.ec.europa.eu
		Organization European Commission Joint Research Centre ralf.steinberger@jrc.ec.europa.eu
Domains	law_politics	
Time coverages	1958-2006	
Geographic coverages	European Union	
Creation	Original source	http://optima.jrc.it/Acquis/
	Creation mode	Mixed
	Creation mode details	Semi-automatic acquisition and processing.
	Creation tools	in-house software
Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Multilinguality type details	The texts were aligned on a sentence level using the statistical aligner Maligna (http://align.sourceforge.net).	
Languages	English	
	Language ID	en
	Size	4200000 tokens
	Polish	
	Language ID	pl
	Size	3800000 tokens
Modality	Modality type	Other

Size	8000000 tokens	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Mixed
	Annotation tool	in-house software
	Start date	2011-08-01
	End date	2011-09-30
	Size	8000000 tokens
	Annotators	Piotr Pęzik
		Position assistant professor
		Contact Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
		Drózdź Łukasz
		Position IT specialist
		Contact Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
		Alignment
	Segmentation level	Sentence

	Format	text/xml	
	Conformance to standards best practices	TEI	
	Annotation mode	Automatic	
	Annotation tool	Maligna	
	Start date	2011-08-01	
	End date	2011-09-30	
	Size	8000000 tokens	
	Annotators	Piotr Pęzik	
		Position	assistant professor
		Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
		Drózd Łukasz	
		Position	IT specialist
		Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
	Domains	law_politics	
	Time coverages	2004-2011	
	Geographic coverages	European Union	
	Creation	Original source	http://europa.eu/rapid/
		Creation mode	Mixed
		Creation mode details	Semi-automatic acquisition and processing.
		Creation tools	in-house software

5.2. PELCRA Polish-English parallel corpora (CC-BY-NC)

General Information

Short name	PELCRA-PAR-2
Description	A subset of the PELCRA Polish parallel corpora licensed under the CC-BY-NC license. This resource contains 257 texts from the PAS Academia journal. Individual headers may override the licensing information.
Identifier	502
Resource type	Corpus
URL	http://pelcra.pl/res/parallel/pelcra-par-2
Version	1.0
Revision	compilation of the corpus
Last update	2011-09-30

Contacts

Piotr Pęzik	
Position	assistant professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Łukasz Drózd	
Position	IT specialist
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use	
IPR holder	Polish Academy of Sciences	
	Short name	PAS

	Department name	Office of Science Promotion
	Contact	PKiN, pl. Defilad 1 00-901 Warsaw academia@pan.pl http://www.academia.pan.pl
Availability start date	2011-11-30	

Licences

CC-BY-NC		
Restrictions of use	Other	
Access medium	Downloadable	
Download location	http://pelcra.pl/res/parallel/pelcra-par-2	
Signatories	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Distribution rights holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Chair of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl

Metadata

Creation date	2011-10-24	
Metadata creators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl

	http://pelcra.pl
Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Drózd Łukasz	
Position	IT specialist
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Source	CESAR
Metadata language name	English
Metadata language ID	en
Metadata last date updated	2013-01-22

Validation

Validated	True	
Type	Formal	
Mode	Automatic	
Details	Validation of the XML structure in accordance with the TEI P5 and XLIFF schemas with custom PELCRA extensions.	
Extent	Full	
Size	710000 tokens	
Tool	xmllint	
Validator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl

Łukasz Dróżdż	
Position	IT specialist
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl

Resource creation

Resource creator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR

	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2011-08-01	
Creation end date	2011-09-30	

Resource documentation

Reports	http://pelcra.pl/projects/documentation/
----------------	---

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Multilinguality type details	The texts were manually aligned on a sentence level using the MemoQ segment alignment tool (http://kilgray.com/products/memoq).	
Languages	Polish	
	Language ID	pl
	Size	323000 tokens
	English	
	Language ID	en
	Size	387000 tokens
Modality	Modality type	Other
Size	710000 tokens	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI

Annotation mode	Mixed	
Annotation tool	MemoQ	
Start date	2011-08-01	
End date	2011-09-30	
Size	710000 tokens	
Annotators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
	Drózd Łukasz	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Alignment		
Segmentation level	Sentence	
Format	text/xml	
Conformance to standards best practices	TEI	
Annotation mode	Manual	
Annotation tool	MemoQ	
Start date	2011-08-01	
End date	2011-09-30	
Size	710000 tokens	
Annotators	Piotr Pęzik	

		Position	assistant professor
		Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
		Drózd Łukasz	
		Position	IT specialist
		Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Domains	science		
Time coverages	2005-2010		
Creation	Original source	http://www.academia.pan.pl/	
	Creation mode	Mixed	
	Creation mode details	Semi-automatic acquisition and processing.	
	Creation tools	in-house software	

5.3. PELCRA Polish spoken corpus (CC-BY-NC)

General Information

Short name	PELCRA-SP-1
Description	A subset of the PELCRA Polish spoken corpus licensed under the CC-BY-NC license. This resource contains 347 transcriptions of recordings made in the years 2000-2010. Individual headers may override the licensing information.
Identifier	503
Resource type	Corpus
URL	http://pelcra.pl/res/spoken/pelcra-sp-1
Version	1.0
Revision	compilation of the corpus
Last update	2011-09-30

Contacts

Piotr Pęzik	
Position	assistant professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Łukasz Drózdź	
Position	IT specialist
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use	
IPR holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Chair of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Availability start date	2011-11-30	

Licences

CC-BY-NC	
Restrictions of use	Other
Access medium	Downloadable

Download location	http://pelcra.pl/res/spoken/pelcra-sp-1	
Signatories	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Distribution rights holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Chair of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl

Metadata

Creation date	2011-10-24	
Metadata creators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
	Drózd Łukasz	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Source	CESAR	

Metadata language name	English
Metadata language ID	en
Metadata last date updated	2013-01-22

Validation

Validated	True	
Type	Formal	
Mode	Automatic	
Details	Validation of the XML structure in accordance with the TEI P5 and XLIFF schemas with custom PELCRA extensions.	
Extent	Full	
Size	1400000 tokens	
Tool	xmllint	
Validator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Drózd	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl

Resource creation

Resource creator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65

	90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Łukasz Drózd	
Position	IT specialist
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
University of Łódź	
Short name	ULodz
Department name	PELCRA group, Department of English Language and Applied Linguistics
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Funding projects	Central and South-East European Resources
Project short name	CESAR
Project ID	271022
URL	http://www.meta-net.eu/projects/cesar/
Funding type	EU funds
Funder	DG INFSO of the European Commission
Country	European Union
Start date	2011-02-01
End date	2013-01-31
Współczesny korpus referencyjny języka polskiego PELCRA	
Project short name	PELCRA
Project ID	2 H01D 008 25
URL	http://pelcra.pl

	Funding type	National funds
	Funder	Ministry of Science and Higher Education
	Country	Poland
	Start date	2003-10-27
	End date	2005-07-25
	Narodowy Korpus Języka Polskiego	
	Project short name	NKJP
	Project ID	R17 003 03
	URL	http://nkjp.pl
	Funding type	National funds
	Funder	Ministry of Science and Higher Education
	Country	Poland
	Start date	2007-12-13
	End date	2011-06-12
Creation start date	2000-01-01	
Creation end date	2011-09-30	

Resource documentation

Reports	http://pelcra.pl/projects/documentation/
----------------	---

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	pl
	Size	1400000 tokens
Modality	Modality type	Other
Size	1400000 tokens	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Utterance
	Format	text/xml
	Conformance to	TEI

	standards best practices	
	Annotation mode	Manual
	Start date	2000-01-01
	End date	2010-12-31
	Size	1400000 tokens
Domains	general	
Time coverages	2000-2010	
Geographic coverages	Poland	
Creation	Creation mode	Manual
	Creation mode details	Manually transcribed recordings.

5.4. ECL Dictionaries

General Information

Short name	ECL Dictionaries
Description	A set of Wikipedia-derived English-Polish and Polish-English thematic dictionaries available for download under the Creative Commons license of potential use in NLP applications. The dictionaries are based on existing Wikipedia categories, but they have also been manually checked for inappropriately-placed entries. The following subjects are covered in this batch of dictionaries: American universities, world cities and villages, Polish artists, Polish journalists, Polish scientists, Polish politicians, Polish companies, Polish catastrophes, Polish media, Polish organizations, Polish universities. The dictionaries are stored in the RDF (Resource Description Framework) format, which is a method for conceptual description or modeling of information that allows storage of additional information, in this case the Wikipedia categories to which the individual entries belong. The categories presented do not reflect the exact Wikipedia structure, but rather conceptual relations between the entries.
Identifier	504
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Terminological resource
URL	http://pelcra.pl/res/ecl-dictionaries
Version	1.0
Last update	2012-06-15

Contacts

Piotr Pęzik

Position	Assistant Professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use	
IPR holder	Piotr Pęzik	
	Position	Assistant Professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Availability start date	2012-06-15	

Licences

CC-BY		
Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://pelcra.pl/res/ecl-dictionaries	
Signatories	Piotr Pęzik	
	Position	Assistant Professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Distribution rights holder	Piotr Pęzik	
	Position	Assistant Professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
User nature	Academic Commercial	

Metadata

Creation date	2012-06-18	
Metadata creators	Maciej Buczek	
	Position	Programmer
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Metadata language name	English	
Metadata language ID	en	
Metadata last date updated	2013-01-22	

Validation

Validated	True	
Type	Formal	
Mode	Manual	
Extent	Full	
Validator	Maciej Buczek	
	Position	Programmer
	Contact	Kościuszki 65

		90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Usage

Foreseen use	NLP applications		
NLP-specific use	Other		
Actual uses	NLP applications		
	NLP-specific use	Natural language understanding	
	Usage project	Central and South-East European Resources	
		Project short name	CESAR
		Project ID	271022
		URL	http://www.meta-net.eu/projects/cesar
		Funding type	EU funds
		Funder	DG INFSO of the European Commission
		Country	European Union
		Start date	2011-02-01
		End date	2013-01-31

Resource creation

Resource creator	Maciej Buczek	
	Position	Programmer
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	University of Łódź	
	Short name	ULodz

	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-04-13	
Creation end date	2012-05-07	

Resource documentation

Reports	http://pelcra.pl/res/eci-dictionaries
Tool documentation type	Online

Lexical conceptual resource

Lexical conceptual resource type	Terminological resource	
Lexical conceptual resource encoding	Encoding level	Semantics
	Linguistic information	Definition/gloss
	Theoretic model	http://www.w3.org/RDF/
Creation	Original source	http://wikipedia.org
	Creation mode	Mixed
	Creation tools	http://www.eclipse.org/

Texts

Media type	text
-------------------	------

Linguality type	Bilingual	
Multilinguality type	Parallel	
Languages	English	
	Language ID	en
	Language script	Latn
	Size	169816 entries
	Polish	
	Language ID	pl
	Language script	Latn
	Size	129889 elements
Size	299705 entries	
Text format	application:rdft+xml	
Character encoding	UTF-8	

5.5. PELCRA EN Lemmatizer

General Information

Short name	PELCRA_EN_Lemmatizer
Description	PELCRA EN Lemmatizer is a British National Corpus-derived lemma dictionary for the Java-based Morfologik stemming library (see http://morfologik.blogspot.com/). It contains a list of unique words appearing in the BNC together with their lemmas and BNC tags that contain part of speech information (see http://www.natcorp.ox.ac.uk/docs/gramtag.html). Note that both the bncLemmatizer.dict and the bncLemmatizer.info files are necessary for the tool to run. Documentation explaining the use of the lemmatizer is available at: http://pelcra.pl/res/en_lemmatizer .
Identifier	505
Resource type	Tool/service
Tool/service type	Other
URL	http://pelcra.pl/res/en_lemmatizer
Version	1.0
Last update	2012-06-15

Contacts

Piotr Pęzik	
Position	Assistant Professor
Contact	Kościuszki 65 90-514 Łódź

	contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use	
IPR holder	Piotr Pęzik	
	Position	Assistant Professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Availability start date	2012-06-15	

Licences

CC-BY		
Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://pelcra.pl/res/en_lemmatizer	
Signatories	Piotr Pęzik	
	Position	Assistant Professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Distribution rights holder	Piotr Pęzik	
	Position	Assistant Professor
	Contact	Kościuszki 65 90-514 Łódź

		contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
User nature	Academic Commercial	

Metadata

Creation date	2012-06-18	
Metadata creators	Maciej Buczek	
	Position	Programmer
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Metadata language name	English	
Metadata language ID	en	
Metadata last date updated	2013-01-22	

Validation

Validated	True	
Type	Formal	
Mode	Manual	
Extent	Full	
Size	31621 kb	
Validator	Maciej Buczek	
	Position	Programmer
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics

	contact@pelcra.pl
--	--

Usage

Resource associated with	http://morfologik.blogspot.com/	
Foreseen use	NLP applications	
NLP-specific use	Derivational morphological analysis	
Actual uses	NLP applications	
	NLP-specific use	Derivational morphological analysis
	Derived resource	http://pelcra.pl/res/parallel/word-aligned/
	Usage project	Central and South-East European Resources
		Project short name CESAR
		Project ID 271022
		URL http://www.meta-net.eu/projects/cesar
		Funding type EU funds
		Funder DG INFSO of the European Commission
		Country European Union
		Start date 2011-02-01
		End date 2013-01-31

Resource creation

Resource creator	Maciej Buczek	
	Position	Programmer
	Contact	Kościuszki 65 90-514 Łódź
		contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65

		90-514 Łódź contact@pelcra.pl http://pelcra.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-04-13	
Creation end date	2012-05-07	

Resource documentation

Reports	http://pelcra.pl/res/en_lemmatizer
Samples location	http://pelcra.pl/res/en_lemmatizer
Tool documentation type	Online

Tool/service

Tool/service type	Other	
Tool/service subtype	dictionary	
Language dependent	True	
Input	English	
	Media type	text
	Language ID	en
	Language variety name	en-gb
	Segmentation level	Word
Output	English	
	Media type	text
	Language ID	en

	Language variety name	en-gb
	Format	tags
	Tagset	http://www.natcorp.ox.ac.uk/docs/c5spec.html
	Segmentation level	Word
Operating system	OS-independent	
Required software	http://morfologik.blogspot.com/	
Required hardware	None	
Required LR	http://sourceforge.net/projects/morfologik/files/morfologik-stemming/	
Running environment details	JRE	
Tool/service creation	Implementation language	Java
	Original source	http://www.natcorp.ox.ac.uk/

5.6. PELCRA Language Detector

General Information

Short name	PELCRA_Language_Detector
Description	The PELCRA language detector is a Java tool for detecting the language of an arbitrary stretch of text developed by the PELCRA team at the University of Łódź, available under the GPL licence. The first version of this tool contains a model for distinguishing between Polish and English. The language detector uses naive Bayes classifier for language detection. The API of the application makes it possible to provide training data for the developments of detectors of other languages.
Identifier	506
Resource type	Tool/service
Tool/service type	Tool
URL	http://pelcra.pl/res/language-detectors
Version	1.0
Last update	2012-07-09

Contacts

Piotr Pęzik	
Position	Assistant Professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl

	http://pelcra.pl
Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use	
IPR holder	Piotr Pęzik	
	Position	Assistant Professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Availability start date	2012-07-09	

Licences

GPL		
Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://pelcra.pl/res/language-detectors	
Signatories	Piotr Pęzik	
	Position	Assistant Professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Distribution rights holder	Piotr Pęzik	
	Position	Assistant Professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl

	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
User nature	Academic Commercial	

Metadata

Creation date	2012-06-18	
Metadata creators	Maciej Buczek	
	Position	Programmer
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Metadata language name	English	
Metadata language ID	en	
Metadata last date updated	2013-01-22	

Usage

Foreseen use	NLP applications		
NLP-specific use	Document classification		
Actual uses	NLP applications		
	NLP-specific use	Document classification	
	Usage project	Central and South-East European Resources	
		Project short name	CESAR
		Project ID	271022
		URL	http://www.meta-net.eu/projects/cesar
		Funding type	EU funds
		Funder	DG INFSO of the European Commission
		Country	European Union

		Start date	2011-02-01
		End date	2013-01-31

Resource creation

Resource creator	Piotr Pęzik	
	Position	Assistant Professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-04-13	
Creation end date	2012-05-07	

Resource documentation

Reports	http://pelcra.pl/res/language-detectors
Samples location	http://pelcra.pl/res/language-detectors

Tool documentation type	Online
--------------------------------	--------

Tool/service

Tool/service type	Tool	
Tool/service subtype	language detector	
Language dependent	True	
Input	English	
	Media type	text
	Language ID	en
	Language variety name	en-gb
	Segmentation level	Other
Operating system	OS-independent	
Required hardware	None	
Required LR	http://sourceforge.net/projects/morfologik/files/morfologik-stemming/	
Running environment details	JRE	
Tool/service creation	Implementation language	Java

5.7. PELCRA Polish-English parallel corpus of literary works (CC-BY)

General Information

Short name	PELCRA-LIT-1
Description	A subset of the PELCRA Polish parallel corpora licensed under the CC-BY license. This resource contains 17 public-domain literary works and their English-Polish/Polish-English translations. The texts have been aligned manually on the sentence level. The texts are provided as TEI P5-compliant XML files with custom PELCRA extensions to mark complex translation equivalence types, and in the XLIFF format.
Identifier	507
Resource type	Corpus
URL	http://pelcra.pl/res/parallel/pelcra-lit-1
Version	1.0
Revision	compilation of the corpus
Last update	2012-06-30

Contacts

Piotr Pęzik	
Position	assistant professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Łukasz Drózd	
Position	IT specialist
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use	
IPR holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Availability start date	2012-06-30	

Licences

CC-BY	
Restrictions of use	Attribution
Access medium	Downloadable
Download location	http://pelcra.pl/res/parallel/pelcra-lit-1

Attribution text	Pęzik P., Ogrodniczuk M., Przepiórkowski A (2011). Parallel and spoken corpora in an open repository of Polish language resources. Human Language Technologies as a Challenge for Computer Science and Linguistics. LTC Poznań 2011.	
Signatories	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Distribution rights holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl

Metadata

Creation date	2012-06-30	
Metadata creators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl

	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Metadata language name	English	
Metadata language ID	en	

Validation

Validated	True	
Type	Formal	
Mode	Automatic	
Details	Validation of the XML structure in accordance with the TEI P5 and XLIFF schemas with custom PELCRA extensions.	
Extent	Full	
Size	3185000 words	
Validation report	Reports	All files are valid XML conforming to the TEI P5 and XLIFF schemas.
Tool	xmllint	
Validator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Resource creation

Resource creator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Drózd	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2011-08-01	

Creation end date	2012-06-30
--------------------------	------------

Resource documentation

Reports	http://pelcra.pl/res/parallel/pelcra-lit-1
----------------	---

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Multilinguality type details	English and Polish original texts with their translations into the respective language, manually aligned on the sentence level.	
Languages	English	
	Language ID	en
	Language script	Latn
	Size	1755000 words
	Polish	
	Language ID	pl
	Language script	Latn
	Size	1430000 words
Modality	Modality type	Written language
	Size	17 texts
Size	17 texts, 3185000 words	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Segmentation	
	Annotation standoff	False
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI_P5
	Annotation mode	Mixed
	Annotation tool	memoQ (http://kilgray.com/products/memoq/)

Start date	2011-08-01	
End date	2012-12-30	
Size	17 texts	
Annotators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Alignment		
Segmentation level	Sentence	
Format	text/xml	
Conformance to standards best practices	TEI	
Annotation mode	Manual	
Annotation tool	memoQ (http://kilgray.com/products/memoq/)	
Start date	2011-08-01	
End date	2012-12-30	
Size	17 texts	
Annotators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź

			contact@pelcra.pl http://pelcra.pl
		Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
		Łukasz Drózdź	
		Position	IT specialist
		Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Domains	literature		
Time coverages	1726-1912		
Creation	Original source	http://gutenberg.org http://wikisource.org http://wolnelektury.pl	
	Creation mode	Manual	
	Creation mode details	The texts were downloaded from public domain repositories. Segmentation and manual alignment were performed using memoQ. Care was taken to represent all non-trivial translation equivalence types.	
	Creation tools	memoQ (http://kilgray.com/products/memoq)	

5.8. PELCRA multilingual parallel corpora (CC-BY)

General Information

Short name	PELCRA-PAR-3
Description	A subset of the PELCRA Polish parallel corpora licensed under the CC-BY license. This resource contains 11300 texts in 6 languages from the CORDIS website, 5556 texts in 28 languages from the RAPID site, 3037 press releases of the European Parliament in 22 languages and 109 press releases of the European Southern Observatory in 17 languages. The texts are sentence-aligned with the mAligna aligner using the Church & Gale algorithm. The texts are provided as TEI P5-compliant XML files with custom PELCRA extensions and in the XLIFF format.
Identifier	508
Resource type	Corpus

URL	http://pelcra.pl/res/parallel/pelcra-par-3
Version	1.0
Revision	compilation of the corpus
Last update	2012-06-30

Contacts

Piotr Pęzik	
Position	assistant professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Łukasz Drózd	
Position	IT specialist
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use	
IPR holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Availability start date	2012-06-30	

Licences

CC-BY		
Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://pelcra.pl/res/parallel/pelcra-par-3	
Attribution text	Pęzik P., Ogrodniczuk M., Przepiórkowski A (2011). Parallel and spoken corpora in an open repository of Polish language resources. Human Language Technologies as a Challenge for Computer Science and Linguistics. LTC Poznań 2011.	
Signatories	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Distribution rights holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl

Metadata

Creation date	2012-06-30	
Metadata creators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Drózdź	
	Position	IT specialist

	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Metadata language name	English	
Metadata language ID	en	

Validation

Validated	True	
Type	Formal	
Mode	Automatic	
Details	Validation of the XML structure in accordance with the TEI P5 and XLIFF schemas with custom PELCRA extensions.	
Extent	Full	
Size	143000000 words	
Validation report	Reports	All files are valid XML conforming to the TEI P5 and XLIFF schemas.
Tool	xmllint	
Validator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź

	PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
--	--

Resource creation

Resource creator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Drózd	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission

	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2011-08-01	
Creation end date	2012-06-30	

Resource documentation

Reports	http://pelcra.pl/res/parallel/pelcra-par-3
----------------	---

Texts

Media type	text	
Linguality type	Multilingual	
Multilinguality type	Parallel	
Multilinguality type details	The texts were aligned on a sentence level using the statistical aligner Maligna (http://align.sourceforge.net).	
Languages	German	
	Language ID	de
	Language script	Latn
	Size	3788000 words
	English	
	Language ID	en
	Language script	Latn
	Size	3907000 words
	Spanish	
	Language ID	es
	Language script	Latn
	Size	4558000 words
	French	
	Language ID	fr
	Language script	Latn
	Size	4456000 words
	Italian	
	Language ID	it
	Language script	Latn

	Size	4247000 words
	Polish	
	Language ID	pl
	Language script	Latn
	Size	3581000 words
Modality	Modality type	Written language
	Size	67787 texts
Size	67787 texts, 24539000 words	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Segmentation	
	Annotation standoff	False
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI_P5
	Annotation mode	Automatic
	Annotation tool	LanguageTool (http://languagetool.org)
	Start date	2011-08-01
	End date	2012-06-30
	Size	67787 texts
	Annotators	Piotr Pęzik
		Position assistant professor
		Contact Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
		Łukasz Dróżdż
		Position IT specialist

	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Alignment		
Segmentation level	Sentence	
Format	text/xml	
Conformance to standards best practices	TEI	
Theoretic model	Church & Gale algorithm (Gale, William A.; Church, Kenneth W. (1993), "A Program for Aligning Sentences in Bilingual Corpora", Computational Linguistics 19 (1): 75–102)	
Annotation mode	Automatic	
Annotation tool	mAligna (http://align.sourceforge.net)	
Start date	2011-08-01	
End date	2012-06-30	
Size	67787 texts	
Annotators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdź	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl

		Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Domains	science		
Time coverages	2003-2012		
Geographic coverages	European Union		
Creation	Original source	http://cordis.europa.eu/news/	
	Creation mode	Mixed	
	Creation mode details	The texts were acquired using a custom-built web crawler. Semi-automatic scripts were used to pipeline text cleanup, segmentation, alignment and import/export procedures.	
	Creation tools	WebLign (http://code.google.com/p/weblign)	
Media type	text		
Linguality type	Multilingual		
Multilinguality type	Parallel		
Multilinguality type details	The texts were aligned on a sentence level using the statistical aligner Maligna (http://align.sourceforge.net).		
Languages	Czech		
	Language ID	cs	
	Language script	Latn	
	Size	107000 words	
	German		
	Language ID	de	
	Language script	Latn	
	Size	125000 words	
	Danish		
	Language ID	da	
	Language script	Latn	
	Size	117000 words	
	English		
	Language ID	en	
	Language script	Latn	
	Size	114000 words	

Spanish	
Language ID	es
Language script	Latn
Size	129000 words
Finnish	
Language ID	fi
Language script	Latn
Size	78000 words
French	
Language ID	fr
Language script	Latn
Size	134000 words
Icelandic	
Language ID	is
Language script	Latn
Size	99000 words
Italian	
Language ID	it
Language script	Latn
Size	119000 words
Dutch	
Language ID	nl
Language script	Latn
Size	115000 words
Norwegian	
Language ID	no
Language script	Latn
Size	115000 words
Polish	
Language ID	pl
Language	Latn

	script	
	Size	104000 words
	Portuguese	
	Language ID	pt
	Language script	Latn
	Size	136000 words
	Russian	
	Language ID	ru
	Language script	Cyrl
	Size	52000 words
	Swedish	
	Language ID	sv
	Language script	Latn
	Size	116000 words
	Turkish	
	Language ID	tr
	Language script	Latn
	Size	83000 words
	Ukrainian	
	Language ID	uk
	Language script	Cyrl
	Size	71000 words
Modality	Modality type	Written language
	Size	1728 texts
Size	1728 texts, 1814000 words	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Segmentation	
	Annotation standoff	False
	Segmentation level	Sentence

Format	text/xml	
Conformance to standards best practices	TEI_P5	
Annotation mode	Automatic	
Annotation tool	LanguageTool (http://languagetool.org)	
Start date	2012-06-01	
End date	2012-06-30	
Size	1728 texts	
Annotators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Alignment		
Segmentation level	Sentence	
Format	text/xml	
Conformance to standards best practices	TEI	
Theoretic model	Church & Gale algorithm (Gale, William A.; Church, Kenneth W. (1993), "A Program for Aligning Sentences in Bilingual Corpora", Computational Linguistics 19 (1): 75–102)	

	Annotation mode	Automatic	
	Annotation tool	mAligna (http://align.sourceforge.net)	
	Start date	2012-06-01	
	End date	2012-06-30	
	Size	1728 texts	
	Annotators	Piotr Pęzik	
		Position	assistant professor
		Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
		Łukasz Dróżdż	
		Position	IT specialist
Contact		Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl	
Organization		University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl	
Domains	science		
Time coverages	2009-2012		
Geographic coverages	European Union		
Creation	Original source	http://www.eso.org	
	Creation mode	Mixed	
	Creation mode details	The texts were acquired using a custom-built web crawler. Semi-automatic scripts were used to pipeline text cleanup, segmentation, alignment and import/export procedures.	
	Creation tools	WebLign (http://code.google.com/p/weblign)	
Media type	text		
Linguality type	Multilingual		
Multilinguality type	Parallel		

Multilinguality type details	The texts were aligned on a sentence level using the statistical aligner Maligna (http://align.sourceforge.net).	
Languages	Bulgarian	
	Language ID	bg
	Language script	Cyrl
	Size	1070000 words
	Czech	
	Language ID	cs
	Language script	Latn
	Size	1401000 words
	Danish	
	Language ID	da
	Language script	Latn
	Size	1256000 words
	German	
	Language ID	de
	Language script	Latn
	Size	1565000 words
	Greek	
	Language ID	el
	Language script	Grek
	Size	1650000 words
	English	
	Language ID	en
	Language script	Latn
	Size	1985000 words
	Spanish	
	Language ID	es
	Language script	Latn
	Size	1911000 words
	Estonian	

Language ID	et
Language script	Latn
Size	987000 words
Finnish	
Language ID	fi
Language script	Latn
Size	1069000 words
French	
Language ID	fr
Language script	Latn
Size	2152000 words
Hungarian	
Language ID	hu
Language script	Latn
Size	1205000 words
Italian	
Language ID	it
Language script	Latn
Size	2127000 words
Lithuanian	
Language ID	lt
Language script	Latn
Size	1118000 words
Latvian	
Language ID	lv
Language script	Latn
Size	1127000 words
Maltese	
Language ID	mt
Language script	Latn

	Size	1134000 words
	Dutch	
	Language ID	nl
	Language script	Latn
	Size	1454000 words
	Polish	
	Language ID	pl
	Language script	Latn
	Size	1514000 words
	Portuguese	
	Language ID	pt
	Language script	Latn
	Size	1725000 words
	Romanian	
	Language ID	ro
	Language script	Latn
	Size	1269000 words
	Slovak	
	Language ID	sk
	Language script	Latn
	Size	1331000 words
	Slovenian	
	Language ID	sl
	Language script	Latn
	Size	1359000 words
	Swedish	
	Language ID	sv
	Language script	Latn
	Size	1403000 words
Modality	Modality type	Written language

	Size	60120 texts
Size	60120 texts, 31810000 words	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Segmentation	
	Annotation standoff	False
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI_P5
	Annotation mode	Automatic
	Annotation tool	LanguageTool (http://languagetool.org)
	Start date	2012-06-01
	End date	2012-06-30
	Size	60120 texts
	Annotators	Piotr Pęzik
		Position assistant professor
		Contact Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
		Łukasz Dróżdź
		Position IT specialist
		Contact Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Alignment	

	Segmentation level	Sentence	
	Format	text/xml	
	Conformance to standards best practices	TEI	
	Theoretic model	Church & Gale algorithm (Gale, William A.; Church, Kenneth W. (1993), "A Program for Aligning Sentences in Bilingual Corpora", Computational Linguistics 19 (1): 75–102)	
	Annotation mode	Automatic	
	Annotation tool	mAligna (http://align.sourceforge.net)	
	Start date	2012-06-01	
	End date	2012-06-30	
	Size	60120 texts	
	Annotators	Piotr Pęzik	
		Position	assistant professor
		Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl	
	Łukasz Dróżdż		
	Position	IT specialist	
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl	
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl	
Domains	law_politics		
Time coverages	2005-2012		
Geographic coverages	European Union		
Creation	Original source	http://www.europarl.europa.eu/	

	Creation mode	Mixed
	Creation mode details	The texts were acquired using a custom-built web crawler. Semi-automatic scripts were used to pipeline text cleanup, segmentation, alignment and import/export procedures.
	Creation tools	WebLign (http://code.google.com/p/weblign)
Media type	text	
Linguality type	Multilingual	
Multilinguality type	Parallel	
Multilinguality type details	The texts were aligned on a sentence level using the statistical aligner Maligna (http://align.sourceforge.net).	
Languages	Arabic	
	Language ID	ar
	Language script	Arab
	Size	1320 words
	Belarussian	
	Language ID	be
	Language script	Cyrl
	Size	311 words
	Bulgarian	
	Language ID	bg
	Language script	Cyrl
	Size	2951000 words
	Czech	
	Language ID	cs
	Language script	Latn
	Size	3519000 words
	Danish	
	Language ID	da
	Language script	Latn
	Size	3582000 words
	German	
	Language ID	de
	Language	Latn

script	
Size	4698000 words
Greek	
Language ID	el
Language script	GreK
Size	4388000 words
English	
Language ID	en
Language script	Latn
Size	4958000 words
Spanish	
Language ID	es
Language script	Latn
Size	5234000 words
Estonian	
Language ID	et
Language script	Latn
Size	2794000 words
Finnish	
Language ID	fi
Language script	Latn
Size	2691000 words
French	
Language ID	fr
Language script	Latn
Size	5627000 words
Irish	
Language ID	ga
Language script	Latn
Size	282000 words
Croatian	

Language ID	hr
Language script	Latn
Size	3300 words
Hungarian	
Language ID	hu
Language script	Latn
Size	3533000 words
Icelandic	
Language ID	is
Language script	Latn
Size	2900 words
Italian	
Language ID	it
Language script	Latn
Size	4790000 words
Lithuanian	
Language ID	lt
Language script	Latn
Size	3069000 words
Latvian	
Language ID	lv
Language script	Latn
Size	2907000 words
Maltese	
Language ID	mt
Language script	Latn
Size	3193000 words
Dutch	
Language ID	nl
Language script	Latn

Size	4229000 words
Norwegian	
Language ID	no
Language script	Latn
Size	6400 words
Polish	
Language ID	pl
Language script	Latn
Size	4533000 words
Portuguese	
Language ID	pt
Language script	Latn
Size	4311000 words
Romanian	
Language ID	ro
Language script	Latn
Size	3196000 words
Russian	
Language ID	ru
Language script	Cyrl
Size	2000 words
Slovak	
Language ID	sk
Language script	Latn
Size	3426000 words
Slovenian	
Language ID	sl
Language script	Latn
Size	3463000 words
Swedish	

	Language ID	sv
	Language script	Latn
	Size	3518000 words
	Turkish	
	Language ID	tr
	Language script	Latn
	Size	5200 words
Modality	Modality type	Written language
	Size	88332 texts
Size	88332 texts, 84910000 words	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Segmentation	
	Annotation standoff	False
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI_P5
	Annotation mode	Automatic
	Annotation tool	LanguageTool (http://languagetool.org)
	Start date	2012-06-01
	End date	2012-06-30
	Size	88332 texts
	Annotators	Piotr Pęzik
		Position assistant professor
		Contact Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Alignment		
Segmentation level	Sentence	
Format	text/xml	
Conformance to standards best practices	TEI	
Theoretic model	Church & Gale algorithm (Gale, William A.; Church, Kenneth W. (1993), "A Program for Aligning Sentences in Bilingual Corpora", Computational Linguistics 19 (1): 75–102)	
Annotation mode	Automatic	
Annotation tool	mAligna (http://align.sourceforge.net)	
Start date	2012-06-01	
End date	2012-06-30	
Size	88332 texts	
Annotators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź

			contact@pelcra.pl http://pelcra.pl
		Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Domains	law_politics		
Time coverages	2004-2012		
Geographic coverages	European Union		
Creation	Original source	http://europa.eu/rapid/	
	Creation mode	Mixed	
	Creation mode details	The texts were acquired using a custom-built web crawler. Semi-automatic scripts were used to pipeline text cleanup, segmentation, alignment and import/export procedures.	
	Creation tools	WebLign (http://code.google.com/p/weblign)	

5.9. OSW Polish-English parallel corpus (CC-BY-NC)

General Information

Short name	PELCRA-PAR-4
Description	A subset of the PELCRA Polish parallel corpora licensed under the CC-BY-NC license. This resource contains 757 Polish-English texts from the Centre for Eastern Studies (OSW) website. The texts are sentence-aligned with the mAligner aligner using the Church & Gale algorithm. The texts are provided as TEI P5-compliant XML files with custom PELCRA extensions and in the XLIFF format.
Identifier	509
Resource type	Corpus
URL	http://pelcra.pl/res/paralle/pelcra-par-4
Version	1.0
Revision	compilation of the corpus
Last update	2012-06-30

Contacts

Łukasz Drózdź	
Position	IT specialist
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl

	http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Piotr Pęzik	
Position	assistant professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use	
IPR holder	Ośrodek Studiów Wschodnich	
	Contact	ul. Koszykowa 6a 00-564 Warszawa info@osw.waw.pl http://www.osw.waw.pl
Availability start date	2012-06-30	

Licences

CC-BY-NC		
Restrictions of use	Academic – non-commercial use Attribution	
Access medium	Downloadable	
Download location	http://pelcra.pl/res/parallel/pelcra-par-4	
Attribution text	Pęzik P., Ogrodniczuk M., Przepiórkowski A (2011). Parallel and spoken corpora in an open repository of Polish language resources. Human Language Technologies as a Challenge for Computer Science and Linguistics. LTC Poznań 2011.	
Signatories	Ośrodek Studiów Wschodnich	
	Contact	ul. Koszykowa 6a 00-564 Warszawa info@osw.waw.pl http://www.osw.waw.pl
Distribution rights holder	Ośrodek Studiów Wschodnich	

	Contact	ul. Koszykowa 6a 00-564 Warszawa info@osw.waw.pl http://www.osw.waw.pl
--	----------------	--

Metadata

Creation date	2012-06-30	
Metadata creators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Drózd	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Metadata language name	English	
Metadata language ID	en	

Validation

Validated	True
Type	Formal
Mode	Automatic
Details	Validation of the XML structure in accordance with the TEI P5 and XLIFF schemas with custom PELCRA extensions.
Extent	Full
Size	1432000 words

Validation report	Reports	All files are valid XML conforming to the TEI P5 and XLIFF schemas.
Tool	xmllint	
Validator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Resource creation

Resource creator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl

	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2011-08-01	
Creation end date	2012-06-30	

Resource documentation

Reports	http://pelcra.pl/res/parallel/pelcra-par-4
----------------	---

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Multilinguality type details	The texts were aligned on a sentence level using the statistical aligner Maligna (http://align.sourceforge.net).	
Languages	English	
	Language ID	en
	Language script	Latn
	Size	796000 words

	Polish	
	Language ID	pl
	Language script	Latn
	Size	635000 words
Modality	Modality type	Written language
	Size	757 texts
Size	757 texts, 1432000 words	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Segmentation	
	Annotation standoff	False
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI_P5
	Annotation mode	Automatic
	Annotation tool	LanguageTool (http://languagetool.org)
	Start date	2011-08-01
	End date	2012-06-30
	Size	757 texts
	Annotators	Piotr Pęzik
		Position assistant professor
		Contact Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
		Organization University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
		Łukasz Dróżdź
		Position IT specialist
		Contact Kościuszki 65 90-514 Łódź

		contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Alignment		
Segmentation level	Sentence	
Format	text/xml	
Conformance to standards best practices	TEI	
Theoretic model	Church & Gale algorithm (Gale, William A.; Church, Kenneth W. (1993), "A Program for Aligning Sentences in Bilingual Corpora", Computational Linguistics 19 (1): 75–102)	
Annotation mode	Automatic	
Annotation tool	mAligna (http://align.sourceforge.net)	
Start date	2011-08-01	
End date	2012-06-30	
Size	757 texts	
Annotators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics

		contact@pelcra.pl
Domains	science	
Time coverages	2003-2012	
Geographic coverages	Europe, Asia	
Creation	Original source	http://www.osw.waw.pl/
	Creation mode	Mixed
	Creation mode details	The texts were acquired using a custom-built web crawler. Semi-automatic scripts were used to pipeline text cleanup, segmentation, alignment and import/export procedures.
	Creation tools	WebLign (http://code.google.com/p/weblign)

5.10. PELCRA time-aligned spoken corpus of Polish (CC-BY-NC)

General Information

Short name	PELCRA-SP-2
Description	A subset of the PELCRA corpus of conversational Polish, time-aligned on the utterance level, licensed under the CC-BY-NC license. This resource contains 386 744 words in 73 transcriptions of over 43 hours of recordings made in the years 2008-2010. The texts are provided as TEI P5-compliant XML files with custom PELCRA extensions and in the XLIFF format.
Identifier	510
Resource type	Corpus
URL	http://pelcra.pl/res/spoken/pelcra-sp-2
Version	1.0
Revision	compilation of the corpus
Last update	2012-06-30

Contacts

Piotr Pęzik	
Position	assistant professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Łukasz Drózdź	

Position	IT specialist
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use	
IPR holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Availability start date	2012-06-30	

Licences

CC-BY-NC		
Restrictions of use	Academic – non-commercial use Attribution Other	
Access medium	Downloadable	
Download location	http://pelcra.pl/res/spoken/pelcra-sp-2	
Attribution text	Pęzik, Piotr. 2012. "Język mówiony w NKJP." In "Narodowy Korpus Języka Polskiego", ed. Adam Przepiórkowski, Mirosław Bańko, Rafał Górski, and Barbara Lewandowska-Tomaszczyk, 37–47. Warszawa: Wydawnictwo Naukowe PWN. http://nkjp.pl/settings/papers/NKJP_ksiazka.pdf	
Signatories	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl

	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Distribution rights holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl

Metadata

Creation date	2012-06-30	
Metadata creators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Drózd	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Metadata language name	English	
Metadata language ID	en	

Validation

--	--

Validated	True	
Type	Formal	
Mode	Automatic	
Details	Validation of the XML structure in accordance with the TEI P5 and XLIFF schemas with custom PELCRA extensions.	
Extent	Full	
Size	368744 words	
Tool	xmllint	
Validator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Drózd	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Resource creation

Resource creator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

	Lukasz Drózdź	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Funding projects	
	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2013-01-31
	Współczesny korpus referencyjny języka polskiego PELCRA	
	Project short name	PELCRA
	Project ID	2 H01D 008 25
	URL	http://pelcra.pl
	Funding type	National funds
	Funder	Ministry of Science and Higher Education
	Country	Poland
	Start date	2003-10-27
	End date	2005-07-25

	Narodowy Korpus Języka Polskiego	
	Project short name	NKJP
	Project ID	R17 003 03
	URL	http://nkjp.pl
	Funding type	National funds
	Funder	Ministry of Science and Higher Education
	Country	Poland
	Start date	2007-12-13
	End date	2011-06-12
Creation start date	2000-01-01	
Creation end date	2011-09-30	

Resource documentation

Reports	http://pelcra.pl/res/spoken
----------------	---

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	pol
	Language script	Latn
	Size	73 texts
Modality	Modality type	Spoken language
	Modality type details	Transcriptions of spontaneous conversations of speakers representing a diverse age (1-90 years) and geographic group.
	Size	73 texts
Size	386744 words, 73 texts, 43 hours	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Speech annotation – orthographic transcription	
	Annotated elements	Background noise Speaker noise
	Annotation standoff	False
	Segmentation	Utterance

level	
Format	text/xml
Conformance to standards best practices	TEI_P5
Annotation mode	Manual
Annotation tool	ELAN (http://www.lat-mpi.eu/tools/elan/)
Start date	2011-05-04
End date	2012-05-12
Size	73 texts
Speech annotation – speaker turns	
Annotation standoff	False
Segmentation level	Utterance
Format	text/xml
Conformance to standards best practices	TEI_P5
Annotation mode	Manual
Annotation tool	ELAN (http://www.lat-mpi.eu/tools/elan/)
Start date	2011-05-04
End date	2012-05-12
Size	73 texts
Speech annotation – sound to text alignment	
Annotation standoff	False
Segmentation level	Utterance
Format	text/xml
Conformance to standards best practices	TEI_P5
Annotation mode	Manual
Annotation mode details	All personal information have been anonymised.
Annotation tool	ELAN (http://www.lat-mpi.eu/tools/elan/)

	Start date	2011-05-04
	End date	2012-05-12
	Size	73 texts
Domains	general	
Time coverages	2008-2010	
Geographic coverages	Poland	
Creation	Creation mode	Manual
	Creation mode details	Recordings of spontaneous conversations manually transcribed orthographically and time-aligned on the utterance level.
	Creation tools	ELAN (http://www.lat-mpi.eu/tools/elan/)

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	pl
	Size	73 files
Modality	Modality type	Spoken language
	Modality type details	Transcriptions of spontaneous conversations of speakers representing a diverse age (1-90 years) and geographic group.
	Size	73 files
Audio size	73 files, 18 gb (43 hours of audio content)	
Audio content	Speech items	Free speech
	Non speech items	Music Noise Sounds
	Noise level	Low
Setting	Naturality	Spontaneous
	Conversational type	Multilogue
	Interactivity	Overlapping
Audio formats	Audio/wav	
	Signal encoding	Linear PCM
	Sampling rate	44100
	Quantization	16
	Compression	False

	Number of tracks	2		
	Recording quality	Medium		
	Size	73 files		
Domains	general			
Time coverages	2008-2010			
Geographic coverages	Poland			
Audio classification	Audio genre	Speech		
	Speech genre	Conversation		
	Register	informal		
	Size	73 files		
Recording	Device	Flash		
	Environment	Other		
	Source channel	Airflow		
	Recorders	University of Łódź		
		Short name	ULodz	
		Department name	PELCRA group, Department of English Language and Applied Linguistics	
		Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl	
Capture	Capturing device type	Microphone		
	Capturing device type details	The conversations were captured using a digital voice recorder.		
	Capturing details	Whenever possible, an attempt was made to take the recordings without the speakers being aware of the fact of being recorded. All participants were asked for permission to use the recordings afterwards.		
	Capturing environment	Complex		
	Person source set	Age range start	1	
		Age range end	2	
Sex of		Mixed		

		persons	
		Origin of persons	Native
		Geographic distribution of persons	Various regions across Poland.
Creation	Creation mode	Manual	
	Creation mode details	Recordings of spontaneous conversations manually transcribed orthographically and time-aligned on the utterance level.	
	Creation tools	ELAN (http://www.lat-mpi.eu/tools/elan/)	

5.11. PELCRA WebLign crawler

General Information

Short name	WEBLIGN
Description	A customizable site-specific crawler for multilingual websites. The tool provides a general crawling infrastructure and several site-specific parsers. The crawling results are stored in a simple relational database (the database schema is provided along with the code.)
Identifier	511
Resource type	Tool/service
Tool/service type	Tool
URL	http://code.google.com/p/weblign/
Version	1.0
Revision	creation of the tool
Last update	2012-06-30

Contacts

Piotr Pęzik	
Position	assistant professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Łukasz Drózd	
Position	IT specialist
Contact	Kościuszki 65

	90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – unrestricted use	
IPR holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Availability start date	2012-06-30	

Licences

BSD-style		
Access medium	Downloadable	
Download location	http://code.google.com/p/weblign/	
Attribution text	http://pelcra.pl	
Signatories	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Distribution rights holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics

	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
--	----------------	---

Metadata

Creation date	2012-06-30	
Metadata creators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Drózd	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Metadata language name	English	
Metadata language ID	en	

Usage

Foreseen use	Human use		
Actual uses	Human use		
	Derived resource	http://pelcra.pl/res/parallel	
	Usage project	Central and South-East European Resources	
		Project short name	CESAR

		Project ID	271022
		URL	http://www.meta-net.eu/projects/cesar/
		Funding type	EU funds
		Funder	DG INFSO of the European Commission
		Country	European Union
		Start date	2011-02-01
		End date	2013-01-31
	Actual use details	The crawler was used to generate selected resources from the PELCRA parallel corpus collection (the ESO, OSW and EuroParl sub-corpora).	

Resource creation

Resource creator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Michał Margielewski	
	Position	developer
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź

	MySQL Server	
Tool/service creation	Implementation language	Java

5.12. PELCRA Word Aligned Corpora

General Information

Short name	PELCRA_WD_ALIGN
Description	A collection of Polish corpora aligned at the word level using the GIZA++ word aligner. Available both in a TEI P5-compliant format and as relational database logical dump. Sentence-level structural annotation is provided as well as alignment confidence scores. Different parts of this resource are available under different licences - please see the appropriate headers for details.
Identifier	512
Resource type	Corpus
URL	http://pelcra.pl/res/parallel/word-aligned
Version	1.0
Last update	2012-07-04

Contacts

Piotr Pęzik	
Position	Assistant Professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use	
IPR holder	Piotr Pęzik	
	Position	Assistant Professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź

	PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Availability start date	2012-06-15

Licences

CC-BY-NC	
Restrictions of use	Academic – non-commercial use Attribution
Access medium	Downloadable
Download location	http://pelcra.pl/res/parallel/word-aligned
Attribution text	Pęzik P., Ogrodniczuk M., Przepiórkowski A (2011). Parallel and spoken corpora in an open repository of Polish language resources. Human Language Technologies as a Challenge for Computer Science and Linguistics. LTC Poznań 2011.
Signatories	Piotr Pęzik
	Position Assistant Professor
	Contact Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Distribution rights holder	Piotr Pęzik
	Position Assistant Professor
	Contact Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
User nature	Academic Commercial

Metadata

Creation date	2012-06-18
Metadata creators	Maciej Buczek
	Position Programmer

	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
Metadata language name	English	
Metadata language ID	en	
Metadata last date updated	2013-01-22	

Validation

Validated	True	
Type	Formal	
Mode	Automatic	
Extent	Full	
Tool	http://linux.about.com/library/cmd/blcmd11_xmllint.htm	
Validator	Piotr Pęzik	
	Position	Assistant Professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl

Usage

Foreseen use	NLP applications
NLP-specific use	Machine translation

Resource creation

Resource creator	Maciej Buczek	
	Position	Programmer
	Contact	Kościuszki 65 90-514 Łódź

		contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Chair of English Language and Applied Linguistics contact@pelcra.pl
	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-06-13	
Creation end date	2012-06-30	

Resource documentation

Reports	http://pelcra.pl/res/parallel/word-aligned
Samples location	http://pelcra.pl/res/parallel/word-aligned
Tool documentation type	Online

Texts

Media type	text
Linguality type	Bilingual
Multilinguality type	Parallel
Multilinguality type details	Word level alignments

Languages	English	
	Language ID	en
	Language script	Latn
	Size	40955095 words
	Polish	
	Language ID	pl
	Language script	Latn
	Size	34416872 words
Modality	Modality type	Written language
Size	77371967 words	
Text format	text/xml	
Character encoding	UTF-8	
Creation	Original source	http://pelcra.pl/res/parallel
	Creation mode	Mixed
	Creation tools	GIZA++

5.13. Spelling and NUMbers Voice database

General Information

Short name	SNUV
Description	SNUV (Spelling and NUMbers Voice database) is a spelling and number and recognition speech database containing over 220 hours of recordings of Polish speakers reading numbers and spelling words, recorded in 22050kHz, 16-bit *.wav files. 210 different participants were paid to produce a sample of their speech through an online spoken data collection platform. Written representation of the recordings is provided with the original sound files. The envisaged application of this resource is to enable the creation of automatic speech recognition (ASR) tools that allow users to spell out words and numbers to be recognized. SNUV has been released under a CC-BY license and can be used for both academic and commercial purposes free of charge.
Identifier	513
Resource type	Corpus
URL	http://pelcra.pl/res/spoken/snuv
Version	1.0
Revision	compilation of the database
Last update	2012-12-07

Contacts

Piotr Pęzik	
Position	assistant professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Łukasz Drózd	
Position	IT specialist
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use	
IPR holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Availability start date	2012-12-07	

Licences

CC-BY		
Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://pelcra.pl/res/spoken/snuv	
Signatories	Piotr Pęzik	
	Position	assistant professor

	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Distribution rights holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl

Metadata

Creation date	2012-06-30	
Metadata creators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Maciej Buczek	
	Position	Programmer
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Metadata language name	English	

Metadata language ID	en
-----------------------------	----

Validation

Validated	True	
Type	Formal	
Mode	Automatic	
Extent	Full	
Size	704625 words	
Tool	xmllint	
Validator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Resource creation

Resource creator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź

		PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Tomasz SzweInik	
	Position	President
	Contact	Zwycięstwa 96/98 81 – 451 Gdynia voicelab@voicelab.pl http://voicelab.pl/
	Organization	VOICE LAB sp. z o.o. voicelab@voicelab.pl
	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Funding projects	
	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2013-01-31
	Central and South-East European Resources	
	Project short name	CESAR_PL
	Project ID	MNiSW 2139/CIP2007-2011/2
	URL	http://en.kpk.gov.pl/index.php?option=com_sobi2&sobi2Task=sobi2Details&catid=56&sobi2Id=218&Itemid=142&lang=pl
	Funding type	National funds
	Funder	Polish Ministry of Science and Higher Education

	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31

Resource documentation

Reports	http://pelcra.pl/res/spoken/snuv
----------------	---

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	pol
	Language script	Latn
Modality	Modality type	Spoken language
	Modality type details	Records of speakers reading out numbers and spelling words.
Size	704625 words, 220 hours	
Character encoding	UTF-8	
Annotation	Speech annotation – sound to text alignment	
	Annotation standoff	False
	Segmentation level	Utterance
	Format	text/plain
	Annotation mode	Manual
	Annotation mode details	All personal information has been anonymised.
	Start date	2012-01-02
	End date	2012-12-06
	Size	704625 words
Domains	general	
Geographic coverages	Poland	
Creation	Creation mode	Manual

	Creation mode details	Recordings of Polish speakers reading out numbers and spellings, time-aligned on the utterance level.
	Creation tools	SNUV Web application (http://snuv.pl)

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	pl
	Size	704625 files
Modality	Modality type	Spoken language
	Modality type details	Recordings of Polish speakers reading out numbers and spellings, time-aligned on the utterance level.
	Size	704625 words
Audio size	704625 words, 34 gb (220 hours of audio content)	
Audio content	Speech items	Isolated digits Isolated words Natural numbers
	Noise level	Low
Setting	Naturality	Prompted
	Conversational type	Monologue
	Interactivity	Other
Audio formats	Audio/wav	
	Signal encoding	Linear PCM
	Sampling rate	22050
	Quantization	16
	Compression	False
	Number of tracks	1
	Recording quality	Medium
	Size	704625 words
Domains	general	
Geographic coverages	Poland	
Audio classification	Audio genre	Speech
	Size	704625 words

Recording	Device	Hard disk	
	Environment	Other	
	Recorders	VOICE LAB sp. z o.o.	
		Short name	VOICE LAB
		Contact	Zwycięstwa 96/98 81 – 451 Gdynia voicelab@voicelab.pl http://voicelab.pl/
Capture	Capturing device type	Close talk microphone	
	Capturing device type details	The utterances were captured using an external microphone.	
	Capturing environment	Complex	
	Person source set	Age range start	11
		Age range end	69
		Sex of persons	Mixed
		Origin of persons	Native
		Geographic distribution of persons	Various regions across Poland.
Creation	Creation mode	Manual	
	Creation mode details	Recordings of Polish speakers reading out prompted numbers and spelling words, time-aligned on the utterance level.	
	Creation tools	SNUV Web application (http://snuv.pl)	

5.14. HASK collocation dictionary (English)

General Information

Short name	HASK-EN-API
Description	An API for the English version of the HASK dictionary of frequent word combinations automatically generated from the British National Corpus. Developed by the PELCRA group at the University of Łódź, HASK dictionaries are essentially phraseological databases meant to be used by linguists, language teachers, lexicographers, language materials developers, translators and other language professionals and casual dictionary users.

Identifier	514
Resource type	Tool/service
Tool/service type	Service
URL	http://pelcra.pl/hask_en
Version	1.0
Revision	public availability of the API
Last update	2012-12-31

Contacts

Piotr Pęzik	
Position	assistant professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use	
IPR holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Availability start date	2012-12-31	

Licences

CC-BY-NC	
Restrictions of use	Academic – non-commercial use Attribution
Execution location	http://pelcra.pl/hask_en
Signatories	Piotr Pęzik

	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Metadata

Creation date	2012-12-31	
Metadata creators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Metadata language name	English	
Metadata language ID	en	

Validation

Validated	True
Type	Formal

Mode	Manual	
Details	The output data was validated as well-formed XML.	
Extent	Full	
Validation report	Reports	The output data is well-formed XML.
Tool	xmllint	
Validator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Resource creation

Resource creator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022

	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2012-12-31
	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	MNiSW 2139/CIP2007-2011/2
	URL	http://en.kpk.gov.pl/index.php?option=com_sobi2&sobi2Task=sobi2Details&catid=56&sobi2Id=218&Itemid=142&lang=pl
	Funding type	National funds
	Funder	Ministry of Science and Higher Education
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-06-01	
Creation end date	2012-12-31	

Resource documentation

Reports	Examples of use are available at http://pelcra.pl/res/hask .
Samples location	http://pelcra.pl/res/hask
Tool documentation type	Online

Tool/service

Tool/service type	Service
Tool/service subtype	API
Language dependent	True
Input	English

	Media type	text
	Language ID	en
Output	English	
	Media type	text
	Language ID	en
Operating system	OS-independent	

5.15. HASK collocation dictionary (Polish)

General Information

Short name	HASK-PL-API
Description	An API for the Polish version of the HASK dictionary of frequent word combinations automatically generated from the Polish National Corpus. Developed by the PELCRA group at the University of Łódź, HASK dictionaries are essentially phraseological databases meant to be used by linguists, language teachers, lexicographers, language materials developers, translators and other language professionals and casual dictionary users.
Identifier	515
Resource type	Tool/service
Tool/service type	Service
URL	http://pelcra.pl/hask_pl
Version	1.0
Revision	public availability of the API
Last update	2012-12-31

Contacts

Piotr Pęzik	
Position	assistant professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use

IPR holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Availability start date	2012-12-31	

Licences

CC-BY-NC		
Restrictions of use	Academic – non-commercial use Attribution	
Execution location	http://pelcra.pl/hask_pl	
Signatories	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Metadata

Creation date	2012-12-31	
Metadata creators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Łukasz Dróżdż	

	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Metadata language name	English	
Metadata language ID	en	

Validation

Validated	True	
Type	Formal	
Mode	Manual	
Details	The output data was validated as well-formed XML.	
Extent	Full	
Validation report	Reports	The output data is well-formed XML.
Tool	xmllint	
Validator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Resource creation

Resource creator	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl

		http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2012-12-31
	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	MNiSW 2139/CIP2007-2011/2
	URL	http://en.kpk.gov.pl/index.php?option=com_sobi2&sobi2Task=sobi2Details&catid=56&sobi2Id=218&Itemid=142&lang=pl
	Funding type	National funds
	Funder	Ministry of Science and Higher Education
	Country	Poland
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-06-01	

Creation end date	2012-12-31
--------------------------	------------

Resource documentation

Reports	Examples of use are available at http://pelcra.pl/res/hask .
Samples location	http://pelcra.pl/res/hask
Tool documentation type	Online

Tool/service

Tool/service type	Service	
Tool/service subtype	API	
Language dependent	True	
Input	Polish	
	Media type	text
	Language ID	pl
Output	Polish	
	Media type	text
	Language ID	pl
Operating system	OS-independent	

5.16. PELCRA Spoken Learner English Corpus

General Information

Short name	PELCRA-PLEC-SP
Description	A subset of the PELCRA PLEC corpus, containing 15 hours (131 000 transcribed words) of recordings of informal interviews with Polish learners of English, time-aligned on the utterance and annotated manually for mispronunciation errors, provided as TEI P5-conformant XML and EAF (ELAN) files.
Identifier	516
Resource type	Corpus
URL	http://pelcra.pl/res/spoken/plec
Version	1.0
Revision	compilation of the corpus
Last update	2012-12-07

Contacts

Piotr Pęzik	
Position	assistant professor
Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use	
IPR holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Availability start date	2012-12-07	

Licences

CC-BY-NC		
Restrictions of use	Attribution	
	Other	
Access medium	Downloadable	
Download location	http://pelcra.pl/res/spoken/plec	
Attribution text	Pęzik, Piotr. 2012. "Towards the PELCRA Learner English Corpus." In "Corpus Data Across Languages and Disciplines", ed. Piotr Pęzik, 28:33–42. Łódź Studies in Language. Peter Lang.	
Signatories	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź

		PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Distribution rights holder	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl

Metadata

Creation date	2012-06-30	
Metadata creators	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Maciej Buczek	
	Position	Programmer
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Metadata language name	English	
Metadata language ID	en	

Validation

Validated	True
------------------	------

Type	Formal	
Mode	Automatic	
Extent	Full	
Size	131542 words	
Tool	xmllint	
Validator	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Resource creation

Resource creator	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Piotr Pęzik	
	Position	assistant professor
	Contact	Kościuszki 65

		90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	University of Łódź	
	Short name	ULodz
	Department name	PELCRA group, Department of English Language and Applied Linguistics
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Country	European Union
	Start date	2011-02-01
	End date	2013-01-31
	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	MNiSW 2139/CIP2007-2011/2
	URL	http://en.kpk.gov.pl/index.php?option=com_sobi2&sobi2Task=sobi2Details&catid=56&sobi2Id=218&Itemid=142&lang=pl
	Funding type	National funds
	Funder	Ministry of Science and Higher Education
	Country	Poland
	Start date	2011-02-01
	End	2013-01-31

date	
Polish Ministry of Science and Higher Education grant	
Project short name	PLEC
Project ID	N N104 205039
URL	http://pelcra.pl/plec/
Funding type	National funds
Funder	Ministry of Science and Higher Education
Country	Poland
Start date	2010-12-01
End date	2013-11-01

Resource documentation

Reports	http://pelcra.pl/res/spoken/plec
----------------	---

Texts

Media type	text	
Linguality type	Monolingual	
Languages	English	
	Language ID	en
	Language script	Latn
Modality	Modality type	Spoken language
	Modality type details	Recordings of interviews with Polish learners of English.
Size	131542 words, 15 hours	
Character encoding	UTF-8	
Annotation	Speech annotation – sound to text alignment	
	Annotated elements	Mispronunciations
	Annotation standoff	True
	Segmentation level	Utterance Word
	Format	text/xml
	Conformance to	TEI_P5

standards best practices	Other
Annotation mode	Manual
Annotation mode details	All personal information has been anonymised.
Start date	2011-07-01
End date	2013-11-01
Size	131542 words
Speech annotation – phonetic transcription	
Annotated elements	Mispronunciations
Annotation standoff	True
Segmentation level	Word
Format	text/xml
Conformance to standards best practices	TEI_P5
Annotation mode	Manual
Annotation mode details	All personal information has been anonymised.
Start date	2011-07-01
End date	2013-11-01
Size	131542 words
Speech annotation – orthographic transcription	
Annotated elements	Mispronunciations
Annotation standoff	True
Segmentation level	Utterance Word
Format	text/xml
Conformance to standards best practices	TEI_P5
Annotation mode	Manual
Annotation	All personal information has been anonymised.

	mode details	
	Start date	2011-07-01
	End date	2013-11-01
	Size	131542 words
Domains	general	
Geographic coverages	Poland	
Creation	Creation mode	Manual
	Creation mode details	Recordings of interviews with Polish learners of English.
	Creation tools	ELAN (http://tla.mpi.nl/tools/tla-tools/elan/)

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	English	
	Language ID	en
	Size	131542 words
Modality	Modality type	Spoken language
	Modality type details	Recordings of interviews with Polish learners of English.
	Size	131542 words
Audio size	131542 words, 8 gb (15 hours of audio content)	
Audio content	Speech items	Free speech
	Non speech items	Noise
	Noise level	Medium
Setting	Naturality	Assisted
	Conversational type	Multilogue
	Interactivity	Overlapping
Audio formats	Audio/wav	
	Signal encoding	Linear PCM
	Sampling rate	44100
	Quantization	16
	Compression	False
	Number of	1

	tracks		
	Recording quality	Medium	
	Size	131542 words	
Domains	general		
Geographic coverages	Poland		
Audio classification	Audio genre	Speech	
	Speech genre	Conversation	
	Register	informal	
	Size	131542 words	
Recording	Device	Hard disk	
	Environment	Other	
	Recorders	University of Łódź	
		Short name	ULodz
		Department name	PELCRA group, Department of English Language and Applied Linguistics
		Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
Capture	Capturing device type	Microphone	
	Capturing device type details	Conversations were captured using an audio console set with external microphones or a voice recorder.	
	Capturing environment	Complex	
	Person source set	Number of persons	119
		Age range start	8
		Age range end	45
		Sex of persons	Mixed
		Origin of persons	Native
Geographic distribution	Łódź region.		

		of persons	
Creation	Creation mode	Manual	
	Creation mode details	Recordings of interviews with Polish learners of English.	
	Creation tools	ELAN (http://tla.mpi.nl/tools/tla-tools/elan/)	

5.17. Polish-Russian Parallel Corpus

General Information

Short name	PELCRA-PAR-5
Description	A manually aligned Polish-Russian parallel corpus of 4 250 000 words from 20 Polish literary works and 1 legal text and 14 Russian literary texts translated into Russian and Polish respectively. The texts were processed and aligned using ABBY Aligner. All corrections to the segmentation and alignment of the translation with the original were performed manually. The texts are provided as TEI P5-compliant XML files with custom PELCRA extensions to mark complex translation equivalence types, and in the XLIFF and TMX formats. The corpus was originally acquired from the University of Warsaw and enhanced by University of Lodz and the Institute of Computer Science, Polish Academy of Sciences.
Identifier	517
Resource type	Corpus
URL	http://pelcra.pl/res/parallel/pelcra-par-5
Version	1.0
Revision	compilation of the corpus
Last update	2012-12-31

Contacts

Adam Przepiórkowski	
Position	Professor, Head of the Linguistic Engineering Group
Contact	Jana Kazimierza 5 01-248 Warsaw adam.przepiorkowski@ipipan.waw.pl http://zil.ipipan.waw.pl/AdamPrzepiorkowski
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Piotr Pęzik	
Position	assistant professor
Contact	Kościuszki 65 90-514 Łódź

	contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl

Distribution

Availability	Available – restricted use	
IPR holder	Uniwersytet Warszawski	
	Contact	Krakowskie Przedmieście 26/28 00-927 Warszawa m.j.lazinski@uw.edu.pl http://uw.edu.pl
Availability start date	2012-12-31	

Licences

CC-BY-NC		
Restrictions of use	Academic – non-commercial use Attribution	
Access medium	Downloadable	
Download location	http://pelcra.pl/res/parallel/pelcra-par-5	
Attribution text	Pęzik P., Ogrodniczuk M., Przepiórkowski A (2011). Parallel and spoken corpora in an open repository of Polish language resources. Human Language Technologies as a Challenge for Computer Science and Linguistics. LTC Poznań 2011.	
Signatories	Uniwersytet Warszawski	
	Contact	Krakowskie Przedmieście 26/28 00-927 Warszawa m.j.lazinski@uw.edu.pl http://uw.edu.pl
Distribution rights holder	Uniwersytet Warszawski	
	Contact	Krakowskie Przedmieście 26/28 00-927 Warszawa m.j.lazinski@uw.edu.pl http://uw.edu.pl

Metadata

Creation date	2012-12-31	
Metadata creators	Łukasz Drózd	

	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Maciej Ogrodniczuk	
	Position	Assistant professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Metadata language name	English	
Metadata language ID	en	

Validation

Validated	True	
Type	Formal	
Mode	Automatic	
Details	Validation of the XML structure in accordance with the TEI P5, XLIFF and TMX schemas with custom extensions.	
Extent	Full	
Validation extent details	validationExtentDetails	
Size	4250000 words	
Validation report	Reports	All files are valid XML conforming to the TEI P5, XLIFF and TMX schemas.
Tool	xmllint	
Validator	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź

	contact@pelcra.pl http://pelcra.pl
Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
Maciej Ogrodniczuk	
Position	Assistant professor
Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl

Resource creation

Resource creator	Łukasz Drózd	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Maciej Ogrodniczuk	
	Position	Assistant professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
	Central and South-East European Resources	
	Project short name	CESAR
Funding projects	Project ID	271022

	URL	http://www.meta-net.eu/projects/cesar/
	Funding type	EU funds
	Funder	DG INFSO of the European Commission
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-07-01	
Creation end date	2012-12-31	

Resource documentation

Reports	http://pelcra.pl/res/parallel/pelcra-par-5
----------------	---

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Multilinguality type details	Polish and Russian original texts with their translations into the respective language, manually aligned on the sentence level.	
Languages	Russian	
	Language ID	ru
	Language script	Cyrl
	Size	2075000 words
	Polish	
	Language ID	pl
	Language script	Latn
	Size	2175000 words
Modality	Modality type	Written language
	Size	35 texts
Size	35 texts, 4250000 words	
Text format	text/xml	
Character encoding	UTF-8	
Annotation	Segmentation	
	Annotation standoff	False
	Segmentation level	Sentence

Format	text/xml	
Conformance to standards best practices	TEI_P5	
Annotation mode	Mixed	
Annotation tool	ABBYY Aligner (http://www.abbyy.com/aligner/)	
Start date	2012-07-01	
End date	2012-12-31	
Size	35 texts	
Annotators	Łukasz Dróżdż	
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Maciej Ogrodniczuk	
	Position	Assistant professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Alignment		
Segmentation level	Sentence	
Format	text/xml	
Conformance to standards best practices	TEI	
Annotation mode	Manual	
Annotation tool	ABBYY Aligner (http://www.abbyy.com/aligner/)	

	Start date	2012-07-01
	End date	2012-12-31
	Size	35 texts
	Annotators	Łukasz Dróżdż
	Position	IT specialist
	Contact	Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl
	Organization	University of Łódź PELCRA group, Department of English Language and Applied Linguistics contact@pelcra.pl
	Annotators	Maciej Ogrodniczuk
	Position	Assistant professor
	Contact	Jana Kazimierza 5 01-248 Warsaw maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group ipi@ipipan.waw.pl
Domains	literature (34 texts) legal (1 texts)	
Time coverages	1844-2000	
Creation	Creation mode	Manual
	Creation mode details	The texts were acquired from rightholders or, where no copyright was applicable, downloaded from public domain repositories. Segmentation and manual alignment were performed using ABBYY Aligner.
	Creation tools	ABBYY Aligner (http://www.abbyy.com/aligner/)

6. UBG resources

6.1. Serbian Wordnet

General Information

Short name	SrpWN
Description	Serbian WordNet (SrpWN) represents a lexical semantic network, containing synsets with

	glosses and various semantic relations, such as antonymy, meronymy, causation, category domain, etc. The initial version of the Serbian Wordnet was produced in the scope of the EU-funded Balkanet project and it contains all synsets from basic concept sets 1 and 2, and two thirds of synsets from basic concept set 3. Through interlingual relations it is connected to English Wordnet (versions 2.0 and 3.0) and wordnets of many other languages. Currently the Serbian Wordnet contains 18,366 synsets (literals 31,274): 1380 adjectives (literals 1887), 2104 verbs (literals 3918), 14,765 nouns (literals 25,298), other 117. 706 synsets are not connected to the PWN, being either Balkan specific concepts (532) or Serbian specific concepts (174). 18,310 synsets have definitions in Serbian, and 1,274 have examples of usage. Semantic relations in SrpWN: hypernym - 16,590; holo_part - 1,298; holo_member - 3,831; holo_portion - 118; near_antonym - 736; be_in_state - 252; causes - 63. From 31,274 literals in SrpWN 10,164 are multi-word units.
Identifier	601
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Wordnet
URL	http://korpus.matf.bg.ac.rs/SrpWN
Version	v3.0_1
Last update	2013-01-31

Contacts

Cvetana Krstev	
Position	Professor
Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Organization	University of Belgrade Faculty of Philology info@fil.bg.ac.rs

Distribution

Availability	Available – restricted use	
IPR holder	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana/
Availability start date	2011-12-01	

Licences

CC-BY-NC	
Restrictions of use	Academic – non-commercial use

Access medium	Downloadable	
Download location	http://korpus.matf.bg.ac.rs/SrpWN	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas/
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas/

Metadata

Creation date	2011-11-17	
Metadata creators	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2013-01-31	

Validation

Validated	True	
Type	Formal	
Size	18366 synsets	
Tool	XMLSpy;VisDic	
Validator	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana

Usage

Access tool	VisDic; LeXimir; VeBranka; Bibliša
Foreseen use	NLP applications
NLP-specific use	Bilingual lexicon induction

	Document classification Information extraction Lexicon acquisition from corpora		
Foreseen use	NLP applications		
NLP-specific use	Document classification Semantic role labelling		
Actual uses	NLP applications		
	NLP-specific use	Information extraction Information retrieval Web services	
	Reports	<p>Ivan Obradović, Cvetana Krstev, Gordana Pavlović-Lažetić, Duško Vitas, “Corpus Based Validation of WordNet Using Frequency Parameters”, in Proceedings of the GWC : Second International WordNet Conference, Brno, Czech Republic, January 20-23, 2004, eds. P. Sojka, K. Pala, P. Smrž, Ch. Fellbaum, P. Vossen, ed. 1, pp. 181-186, Masaryk University, Brno, 2004.</p> <p>Svetla Koeva, Cvetana Krstev, Duško Vitas, “Morpho-semantic Relations in WordNet - a Case Study for two Slavic Languages”, In the Proceedings of Global WordNet Conference 2008, eds. Attila Tanacs et al, University of Szeged, Department of Informatics, pp. 239-253, 2008.</p> <p>Cvetana Krstev, Ivan Obradović, Duško Vitas, “An Approach to the Development of Language Specific Concepts in Wordnets”, In Southern Journal of Linguistics, Special Theme: South Slavic and Balkan Languages, Mila Dimitrova-Vulchanova (ed.), Vo. 29, No. 1/2, pp. 106-118, Department of Modern Linguistics, University of Mississippi, 2008.</p> <p>C. Krstev and B. Djordjević and S. Antonić and N. Ivković-Berček and Z. Zorica and V. Crnogorac and L. Macura, "Cooperative Work in Further Development of Serbian Wordnet," INFOtheca, vol. 9, pp. 59a-78a, May 2008.</p>	
	Usage project	Serbian Language and its Resources: Theory, Description and Applications	
		Project ID	ON 178006
		Funding type	National funds
		Funder	Serbian Ministry of Education and Science
Country		Serbia	
Start date		2011-01-01	
End date	2015-12-31		

Resource creation

Resource creator	Cvetana Krstev	
	Position	Professor

	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
	Gordana Pavlović-Lažetić	
	Position	Professor
	Contact	gordana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~gordana/
	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
	Ivan Obradović	
	Position	Professor
	Contact	ivano@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/LicnePrezentacije/ivan_obradovic/
	Organization	University of Belgrade Faculty of Geology and Mining sladja@rgf.bg.ac.rs
Funding projects	Design and Development of a Multilingual Balkan WordNet	
	Project short name	BalkaNet
	Project ID	IST-2000-29388
	Funding type	EU funds
	Start date	2002-01-01
	End date	2004-12-31
Creation start date	2002-06-01	

Resource documentation

Tool documentation type	Online
--------------------------------	--------

Lexical conceptual resource

Lexical conceptual resource type	Wordnet	
Lexical conceptual resource encoding	Encoding level	Semantics
	Linguistic information	Part of speech Semantics – relations

	Conformance to standards best practices	Word net
--	--	----------

Texts

Media type	text		
Linguality type	Monolingual		
Languages	Serbian		
	Language ID	srp	
	Language script	Latin	
	Language variety	Language variety type	Dialect
		Language variety name	Ekavian
		Size	18366 synsets
Modality	Modality type	Written language	
Size	18366 synsets		
Character encoding	UTF-8		

6.2. Corpus of Contemporary Serbian

General Information

Short name	SrpKor
Description	The Corpus of contemporary Serbian, SrpKor, consists of 4,925 texts. Total size of SrpKor is 118,767,279 words. It is lemmatized and PoS tagged using TreeTagger. SrpKor texts consist of: fiction written by Serbian authors in 20th and 21th century (10,191,092 words), various scientific texts from various domains (both humanities and sciences) (3,542,169 words), legislative texts (6,874,318 words) and general texts (98,159,700 words). General texts represent daily news published in newspaper "Politika" 2000-2002 and 2005-2010, texts in journals and magazines 1991-2002 ("Danica", "Ebit", "Ekonomist", "Glasnik", "NIN", "Ilustrovana politika", "Kalibar", "Moje srce", "Mostovi", "Pravoslavlje", "Svet", "Teološki pogledi", "Trn", "Viva", "Republika"), internet portal texts 2011-2012 (Peščanik), TANJUG agency news 1995-96, newspaper feuilletons published in newspapers "Politika" (2001-2003), "Večernje novosti" (2008-2011) and "Danas" (2002-2006).
Identifier	602
Resource type	Corpus
URL	http://www.korpus.matf.bg.ac.rs
Version	v2.1

Last update	2012-08-01
--------------------	------------

Contacts

Duško Vitas	
Position	Professor
Contact	vas@matf.bg.ac.rs http://www.matf.bg.ac.rs/~vas
Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Distribution

Availability	Available – restricted use	
IPR holder	Duško Vitas	
	Position	Professor
	Contact	vas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Availability start date	2011-12-01	

Licences

CC-BY-NC		
Restrictions of use	Academic – non-commercial use	
Access medium	Accessible through interface	
Execution location	http://www.korpus.matf.bg.ac.rs	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vas@matf.bg.ac.rs

		http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Metadata

Creation date	2011-11-17	
Metadata creators	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
	Miloš Utvić	
	Position	Teaching Assistant
	Contact	misko@matf.bg.ac.rs
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2012-07-27	

Validation

Validated	True	
Type	Formal	
Mode	Manual	
Extent	Full	
Validator	Miloš Utvić	
	Position	Teaching Assistant
	Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs
	Organization	University of Belgrade Faculty of Philology info@fil.bg.ac.rs

Usage

Access tool	Corpus query processor (CQP)
Resource associated with	IMS Open Corpus Workbench (CWB), http://cwb.sourceforge.net
Foreseen use	NLP applications
NLP-specific use	Annotation

	Lexicon acquisition from corpora Morphological analysis Morphosyntactic tagging	
Actual uses	NLP applications	
	NLP-specific use	Annotation
	Reports	Cvetana Krstev, Duško Vitas, "Corpus and Lexicon - Mutual Incompleteness", in Proceedings of the Corpus Linguistics Conference, 14-17 July 2005, Birmingham, eds. Pernilla Danielsson and Martijn Wagenmakers, ISSN 1747-9398 Duško Vitas, Cvetana Krstev, "Processing of Corpora of Serbian Using Electronic Dictionaries", in Prace Filologiczne, 2012 (to appear)

Resource creation

Resource creator	Miloš Utvić	
	Position	Teaching Assistant
	Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs
	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
Funding projects	Serbian Language and its Resources: Theory, Description and Applications	
	Project ID	ON 178006
	Funding type	National funds
	Funder	Serbian Ministry of Education and Science
	Country	Serbia
	Start date	2011-01-01
	End date	2015-12-31
Creation start date	2011-12-01	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Serbian	
	Language ID	srp
	Language script	Latin
	Size	118767279 words

	Language variety	Language variety type	Dialect
		Language variety name	Ekavian
		Size	118767279 words
Modality	Modality type	Written language	
Size	118767279 words		
Character encoding	ISO-8859-1		
Annotation	Lemmatization		
	Annotation standoff	True	
	Segmentation level	Word	
	Format	text/plain	
	Tagset	http://www.korpus.matf.bg.ac.rs/SrpLemKor/tagset.html	
	Conformance to standards best practices	Other	
	Annotation mode	Automatic	
	Annotation tool	Tree Tagger	
	Start date	2011-10-01	
	Size	118767279 words	
	Annotators	Miloš Utvić	
		Position	Teaching Assistant
Contact		misko@matf.bg.ac.rs http://www.fil.bg.ac.rs	
Domains	literature (10191092 words) science (3542169 words) law_politics (6874318 words) general (98159700 words)		
Creation	Original source	downloading from Web; retyping; scanning and OCR; obtaining from authors and translators	
	Creation mode	Automatic	

6.3. Serbian Lemmatized and PoS Annotated Corpus

General Information

Short name	SrpLemKor
Description	The Serbian Lematized and PoS Annotated Corpus consists of a sample of various texts from SrpKor. It is lemmatized and PoS tagged using TreeTagger. It consists of: daily news published in newspaper "Politika" in december 2009 (1,002,739 words), newspaper feuilletons (1,010,676) published in newspapers "Politika" (2001-2003) and "Danas" (2002-2006), fiction written by Serbian authors in 20th century (869,445), various scientific texts from various domains (both humanities and sciences) (773,119), and legislative texts (107,373). Total size of corpus is 3,763,352 words. More about the content of this corpus can be found at: http://www.korpus.matf.bg.ac.rs/SrpLemKor/SrpLemKor_2011_11.pdf
Identifier	603
Resource type	Corpus
URL	http://www.korpus.matf.bg.ac.rs/SrpLemKor
Version	v1.0
Last update	2011-12-01

Contacts

Miloš Utvić	
Position	Teaching Assistant
Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs
Organization	University of Belgrade Faculty of Philology info@fil.bg.ac.rs

Distribution

Availability	Available – unrestricted use	
IPR holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Availability start date	2011-12-01	

Licences

CC-BY-NC	
Restrictions of use	Academic – non-commercial use
Access medium	Downloadable

Download location	http://www.korpus.matf.bg.ac.rs/SrpLemKor	
Execution location	http://www.korpus.matf.bg.ac.rs/SrpLemKor	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Metadata

Creation date	2011-11-17	
Metadata creators	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2011-11-17	

Validation

Validated	True	
Type	Formal	
Mode	Manual	
Extent	Full	
Validator	Miloš Utvić	
	Position	Teaching Assistant
	Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs

Usage

Access tool	Corpus query processor (CQP)	
Resource associated with	IMS Open Corpus Workbench (CWB), http://cwb.sourceforge.net	
Foreseen use	NLP applications	
NLP-specific use	Annotation Lexicon acquisition from corpora Morphological analysis Morphosyntactic tagging	
Actual uses	NLP applications	
	NLP-specific use	Annotation
	Reports	Zoran Popović, Taggers applied on texts in Serbian, Infotheca, Vol. XI (2), Belgrade, 2010

Resource creation

Resource creator	Miloš Utvić	
	Position	Teaching Assistant
	Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs
	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
Funding projects	Serbian Language and its Resources: Theory, Description and Applications	
	Project ID	ON 178006
	Funding type	National funds
	Funder	Serbian Ministry of Education and Science
	Country	Serbia
	Start date	2011-01-01
	End date	2015-12-31
Creation start date	2011-12-01	

Texts

Media type	text
Linguality type	Monolingual
Languages	Serbian

	Language ID	srp	
	Language script	Latin	
	Size	3763352 words	
	Language variety	Language variety type	Dialect
		Language variety name	Ekavian
		Size	3763352 words
Modality	Modality type	Written language	
Size	3763352 words		
Character encoding	ISO-8859-1		
Annotation	Lemmatization		
	Annotation standoff	True	
	Segmentation level	Word	
	Format	text/plain	
	Tagset	http://www.korpus.matf.bg.ac.rs/SrpLemKor/tagset.html	
	Conformance to standards best practices	Other	
	Annotation mode	Automatic	
	Annotation tool	TreeTagger	
	Start date	2011-10-01	
	Size	3763352 words	
	Annotators	Miloš Utvić	
		Position	Teaching Assistant
Contact		misko@matf.bg.ac.rs http://www.fil.bg.ac.rs	
Domains	literature (869445 words) science (773119 words) law_politics (107373 words) general (2013415 words)		
Creation	Original source	downloading from Web; retyping	
	Creation mode	Automatic	

6.4. French-Serbian Aligned Corpus

General Information

Short name	SrpFranKor
Description	The corpus includes French or Serbian source literary and newspaper texts and their translations. The alignment was performed on the subsentential level. Texts are segment and aligned automatically and then manually checked to obtain one-to-one alignment (in most of the cases). The corpus contains 32 literary texts: 29 French originals with Serbian translation (one with two translations), 2 Serbian originals with French translations, and one English novel translated to French and Serbian. The corpus also contains all articles from the issue of "Le monde diplomatique" from May 2001. The size of the corpus is 59,425 aligned segments and 1,948,679 words (1,063,564 in the French part, 885,115 in the Serbian part). More about the content of this corpus can be found at: http://www.korpus.matf.bg.ac.rs/SrpFranKor/SrpFranKor_2013_01.pdf
Identifier	604
Resource type	Corpus
URL	http://www.korpus.matf.bg.ac.rs/SrpFranKor
Version	v2.2
Last update	2013-01-31

Contacts

Duško Vitas	
Position	Professor
Contact	vitas@matf.bg.ac.rs http://www.matf.bg.ac.rs/~vitas
Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Distribution

Availability	Available – restricted use	
IPR holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Availability start date	2011-12-01	

Licences

CC-BY-NC		
Restrictions of use	Academic – non-commercial use	
Access medium	Accessible through interface	
Execution location	http://www.korpus.matf.bg.ac.rs/SrpFranKor	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Metadata

Creation date	2011-11-18	
Metadata creators	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2013-01-31	

Validation

Validated	True
Type	Formal
Mode	Mixed
Extent	Full

Size	61 files	
Validator	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Usage

Access tool	Corpus query processor (CQP)	
Resource associated with	IMS Open Corpus Workbench (CWB), http://cwb.sourceforge.net	
Foreseen use	Human use NLP applications	
NLP-specific use	Lexicon acquisition from corpora	
Actual uses	Human use	
	NLP-specific use	Bilingual lexicon induction Lexicon acquisition from corpora Temporal expression recognition Terminology extraction
	Reports	Duško Vitas, Cvetana Krstev, "Literature and Aligned Texts", in Readings in Multilinguality, eds. Milena Slavcheva, Galia Angelova and Kiril Simov, pp. 148-155, Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria, 2006. Duško Vitas, Cvetana Krstev, Eric Laporte, "Preparation and exploitation of Bilingual Texts", in Lux Coreana, No. 1, pp. 110-132, Han-Seine, 2006. Duško Vitas, Cvetana Krstev, "Structural derivation and meaning extraction: a comparative study on French-Serbo-Croatian parallel texts", in Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora, eds. Geoff Barnbrook, Pernilla Danielsson, Michaela Mahlberg, pp. 166-178, The University of Birmingham Press, Birmingham, 2005.

Resource creation

Resource creator	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
Funding projects	Serbian Language and its Resources: Theory, Description and Applications	
	Project ID	ON 178006

	Funding type	National funds
	Funder	Serbian Ministry of Education and Science
	Country	Serbia
	Start date	2011-01-01
	End date	2015-12-31
Creation start date	1999-07-01	
Creation end date	2012-08-01	

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Languages	Serbian	
	Language ID	srp
	Language script	Latin
	Size	885115 words
	French	
	Language ID	fra
	Language script	Latin
	Size	1063564 words
Size	1948679 words	
Character encoding	UTF-8	
Annotation	Alignment	
	Annotation standoff	True
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Mixed
	Start date	1999-07-01
	End date	2013-01-31

	Annotators	Duško Vitas	
		Position	Professor
		Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
Domains	general (166597 words) literature (1782082 words)		
Creation	Original source	downloading from Web; retyping; scanning and OCR; obtaining from authors and translators	
	Creation mode	Mixed	

6.5. Multilingual Edition of Verne's Novel "Around the World in 80 Days"

General Information

Short name	Verne80days
Description	This edition contains 18 editions of Jules Verne's novel "Around the World in 80 Days" - French original and 17 translations. The alignment was performed on the subsentential level. Texts are segment and aligned automatically and then manually checked to obtain one-to-one alignment (in most of the cases). In this edition all translations are aligned with either French, English or Serbian version, while all languages represented in CESAR project are aligned with each other. There is a total of 32 aligned texts. All aligned texts are in TMX format and HTML format for visualization. More about the content of this corpus can be found at: http://www.korpus.matf.bg.ac.rs/Verne80days/Verne80days_2012_07.pdf
Identifier	605
Resource type	Corpus
URL	http://www.korpus.matf.bg.ac.rs/Verne80days
Version	v2.1
Last update	2012-07-08

Contacts

Duško Vitas	
Position	Professor
Contact	vitas@matf.bg.ac.rs http://www.matf.bg.ac.rs/~vitas
Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Distribution

--	--

Availability	Available – restricted use	
IPR holder	Duško Vitas	
	Position	Professor
	Contact	vas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Availability start date	2011-12-01	

Licences

CC-BY-NC		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://www.korpus.matf.bg.ac.rs/Verne80days	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Metadata

Creation date	2011-11-18	
Metadata creators	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Source	CESAR	

Metadata language ID	en-us
Metadata last date updated	2013-01-31

Validation

Validated	True	
Type	Formal	
Mode	Mixed	
Extent	Full	
Size	32 files	
Validator	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Usage

Access tool	Corpus query processor (CQP)	
Resource associated with	IMS Open Corpus Workbench (CWB), http://cwb.sourceforge.net	
Foreseen use	Human use NLP applications	
NLP-specific use	Lexicon acquisition from corpora	
Actual uses	Human use	
	NLP-specific use	Bilingual lexicon induction Lexicon acquisition from corpora Temporal expression recognition Terminology extraction
	Reports	Duško Vitas, Svetla Koeva, Cvetana Krstev, Ivan Obradović, “Tour du monde through the dictionaries”, Actes du 27 ^{eme} Colloque International sur le Lexique et la Gamme, L'Aquila, 10-13 septembre 2008, eds. M. Constant, T. Nakamura, M. De Gioia, S. Vecchiato, pp.249-256, Universite Paris-Est, Institut Gaspard-Monge, 2008 Emeline Lecuit, Denis Maurel, Duško Vitas, Cvetana Krstev, “Temporal Expressions: Comparisons in a Multilingual Corpus”, in Proceedings of 4th Language & Technology Conference, November 6-8, 2009, Poznań, Poland, ed. Zygmunt Vetulani, IMPRESJA Wydawnictwa Elektroniczne S.A., Poznań, 2009

Resource creation

Resource creator	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
Funding projects	Serbian Language and its Resources: Theory, Description and Applications	
	Project ID	ON 178006
	Funding type	National funds
	Funder	Serbian Ministry of Education and Science
	Country	Serbia
	Start date	2011-01-01
	End date	2015-12-31
Creation start date	1998-07-01	
Creation end date	2012-07-08	

Texts

Media type	text	
Linguality type	Multilingual	
Multilinguality type	Parallel	
Languages	Serbian	
	Language ID	srp
	Language script	Latin
	Size	58676 words
	French	
	Language ID	fra
	Language script	Latin
	Size	71687 words
	Bulgarian	
	Language ID	bul
	Language script	Cyrillic
	Size	58678 words
	Croatian	
	Language ID	hrv

Language script	Latin
Size	58772 words
Macedonian	
Language ID	mac
Language script	Cyrillic
Size	77255 words
Slovenian	
Language ID	slv
Language script	Latin
Size	62945 words
Polish	
Language ID	pol
Language script	Latin
Size	66277 words
English	
Language ID	eng
Language script	Latin
Size	67947 words
German	
Language ID	ger
Language script	Latin
Size	63496 words
Spanish	
Language ID	spa
Language script	Latin
Size	65502 words
Portuguese	
Language ID	por
Language script	Latin
Size	65012 words

	Greek	
	Language ID	gre
	Language script	Greek
	Size	68063 words
	Italian	
	Language ID	ita
	Language script	Latin
	Size	63976 words
	Dutch	
	Language ID	dut
	Language script	Latin
	Size	70372 words
	Hungarian	
	Language ID	hun
	Language script	Latin
	Size	55816 words
	Albanian	
	Language ID	alb
	Language script	Latin
	Size	67686 words
	Chinese	
	Language ID	chi
	Language script	Simplified Chinese
	Size	116566 words
	Slovak	
	Language ID	slo
	Language script	Latin
	Size	57113 words
Size	1215839 words	
Character encoding	UTF-8	
Annotation	Alignment	

	Annotation standoff	True
	Segmentation level	Sentence
	Format	text/xml
	Conformance to standards best practices	TEI
	Annotation mode	Mixed
	Start date	1998-07-01
	End date	2012-07-08
	Annotators	Duško Vitas
		Position Professor
		Contact vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
Domains	literature	
Creation	Original source	downloading from Web; retyping; scanning and OCR; obtaining from authors and translators
	Creation mode	Mixed

6.6. Organizing digitized material

General Information

Short name	InfoBeaver
Description	This tool is an application for collecting and presenting multimedia informations. It works with multimedial documents and enables database search using different criteria. For various multimedia documents metadata describing them as well as links to their location (on web or locally) are stored into NDX database. Metadata define search criteria that is enabled through web interface. The demo-version illustrates its functionalities with soem data about CESAR project and its participants.
Identifier	606
Resource type	Tool/service
Tool/service type	Other
URL	http://cesar.matf.bg.ac.rs/
Version	v1.0
Last update	2011-12-01

Contacts

Ivana Tanasijević

Position	Assistant
Contact	ivana@math.rs http://www.math.rs/~ivana
Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Distribution

Availability	Available – restricted use	
IPR holder	Ivana Tanasijević	
	Position	Assistant
	Contact	ivana@math.rs http://www.math.rs/~ivana
Availability start date	2011-12-01	

Licences

GPL		
Restrictions of use	Academic – non-commercial use	
Access medium	Web executable	
Execution location	http://www.korpus.matf.bg.ac.rs/InfoBeaver/	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Metadata

Creation date	2011-11-28

Metadata creators	Ivana Tanasijević	
	Position	Assistant
	Contact	ivana@math.rs http://www.math.rs/~ivana
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2011-11-28	

Validation

Validated	True	
Mode	Manual	
Validator	Ivana Tanasijević	
	Position	Assistant
	Contact	ivana@math.rs http://www.math.rs/~ivana

Usage

Actual uses	Human use	
	Reports	<p>Ivana Tanasijević, Biljana Sikimić, Staša Vujičić Stanković, Digitizing and organizing multimedia collection of cultural heritage of the Balkans, National Center for Digitalization (NCD), The X National Conference New Technologies and Standards: Digitization of National Heritage, 33-23.9.2011, Faculty of Mathematics, Belgrade</p> <p>Ivana Tanasijević, Digital tourist map of Belgrade, National Center for Digitalization (NCD), The X National Conference New Technologies and Standards: Digitization of National Heritage, 22-23.9.2011, Faculty of Mathematics, Belgrade</p>

Resource creation

Resource creator	Ivana Tanasijević	
	Position	Assistant
	Contact	ivana@math.rs http://www.math.rs/~ivana
Funding projects	Infrastructure for E-Learning in Serbia	
	Project ID	III 47003
	Funding type	National funds
	Funder	Serbian Ministry of Education and Science
	Country	Serbia

	Start date	2011-01-01
	End date	2015-12-31
Creation start date	2011-06-01	

Resource documentation

Samples location	http://www.korpus.matf.bg.ac.rs/InfoBeaver/demo/
Tool documentation type	None

Tool/service

Tool/service type	Other			
Tool/service subtype	organizing digitized material			
Language dependent	False			
Input	Media type	audio image text video		
Output	Media type	audio image text video		
Operating system	Linux			
Required software	NXDB eXist			
Required hardware	None			
Required LR s	none			
Tool/service evaluation	Evaluated	True		
	Level	Diagnostic		
	Evaluators	Ivana Tanasije vić		
		Position	Assistant	
		Contact	ivana@math.rs http://www.math.rs/~ivana	

6.7. English-Serbian Aligned Corpus

General Information

Short name	SrpEngKor
Description	This corpus consists of English source texts translated into Serbian, and Serbian source

	texts translated into English, and several aligned English and Serbian translations of literary texts originally in French. The texts belong to various domains: fiction, general news, scientific journals, web journalism, health, law, education, movie sub-titles. The corpus also contains several Serbian translations of texts from the 'Acquis communautaire' corpus and from the 'Intera' corpus aligned with their originals. The alignment was performed on the subsentential level. The texts were segmented and aligned automatically and then manually checked. In most cases the alignment is one-to-one. The size of the corpus is 5,078,280 words (2,672,911 in the English part, 2,405,369 in the Serbian part). More about the content of this corpus can be found at: http://www.korpus.matf.bg.ac.rs/SrpEngKor/SrpEngKor_2013_01.pdf
Identifier	607
Resource type	Corpus
URL	http://www.korpus.matf.bg.ac.rs/SrpEngKor
Version	v1.1
Last update	2013-01-01

Contacts

Duško Vitas	
Position	Professor
Contact	vitas@matf.bg.ac.rs http://www.matf.bg.ac.rs/~vitas
Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Distribution

Availability	Available – restricted use	
IPR holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Availability start date	2012-08-01	

Licences

CC-BY-NC	
Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface

Execution location	http://www.korpus.matf.bg.ac.rs/SrpEngKor	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Metadata

Creation date	2012-07-27	
Metadata creators	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2013-01-01	

Validation

Validated	True	
Type	Formal	
Mode	Mixed	
Extent	Full	
Size	62 files	
Validator	Miloš Utvić	
	Position	Teaching Assistant
	Contact	misko@matf.bg.ac.rs

	Organization	University of Belgrade Faculty of Philology info@fil.bg.ac.rs
--	---------------------	--

Usage

Access tool	Corpus query processor (CQP)	
Resource associated with	IMS Open Corpus Workbench (CWB), http://cwb.sourceforge.net	
Foreseen use	Human use NLP applications	
NLP-specific use	Lexicon acquisition from corpora	
Actual uses	Human use	
	NLP-specific use	Bilingual lexicon induction Lexicon acquisition from corpora Temporal expression recognition Terminology extraction
	Reports	Duško Vitas, Cvetana Krstev, "Structural derivation and meaning extraction: a comparative study on French-Serbo-Croatian parallel texts", in Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora, eds. Geoff Barnbrook, Pernilla Danielsson, Michaela Mahlberg, pp. 166-178, The University of Birmingham Press, Birmingham, 2005. Duško Vitas, Cvetana Krstev, "Construction and Exploitation of X-Serbian Bitexts", in Multilingual Processing in Eastern and Southern EU Languages: Low-Resourced Technologies and Translation, eds. Cristina Vertran and Walther v. Hahn, pp. 206-226, Cambridge Scholar Publishing, Newcastle upon Tyne, 2012.

Resource creation

Resource creator	Ivan Obradović	
	Position	Professor
	Contact	ivano@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/LicnePrezentacije/ivan_obradovic/
	Organization	University of Belgrade Faculty of Geology and Mining sladja@rgf.bg.ac.rs
	Miloš Utvić	
	Position	Teaching Assistant
	Contact	misko@matf.bg.ac.rs
	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs

	http://poincare.matf.bg.ac.rs/~vitas
	Cvetana Krstev
Position	Professor
Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana/
Funding projects	Serbian Language and its Resources: Theory, Description and Applications
	Project ID ON 178006
	Funding type National funds
	Funder Serbian Ministry of Education and Science
	Country Serbia
	Start date 2011-01-01
	End date 2015-12-31
Creation start date	1999-07-01
Creation end date	2013-01-01

Texts

Media type	text
Linguality type	Bilingual
Multilinguality type	Parallel
Languages	Serbian
	Language ID srp
	Language script Latin
	Size 2405369 words
	English
	Language ID eng
	Language script Latin
	Size 2672991 words
Size	5078280 words
Character encoding	UTF-8
Annotation	Alignment
	Annotation standoff True
	Segmentation level Sentence
	Format text/xml

	Conformance to standards best practices	TEI	
	Annotation mode	Mixed	
	Start date	1999-07-01	
	End date	2013-01-01	
	Annotators	Miloš Utvić	
		Position	Teaching Assistant
		Contact	misko@matf.bg.ac.rs
Domains	general (909079 words) literature (1364655 words) science (729947 words) law_politics (2074599 words)		
Creation	Original source	downloading from Web; retyping; scanning and OCR; obtaining from authors and translators	
	Creation mode	Mixed	

6.8. Serbian NooJ module

General Information

Short name	SrpNooJ
Description	<p>Serbian NooJ module (SrpNooJ) was produced in the scope of the EU-funded CESAR project. It consists of a set of resources in both alphabets that are in use for Serbian: Cyrillic and Latin. Each set consists of: the dictionary properties' definition file (metadata), one text – a novel “Dva carstva” (Two empires) from a Serbian author Branimir Ćosić comprising of 106684 tokens, a sample dictionary in readable form with 35 lemma that belong to 9 grammatical classes, with examples of multiword units and derivational morphology, a sample of morphological grammars used for lemmas from a sample dictionary – three for simple nouns, two for adjectives, two for verbs, and one for a multiunit noun, a readable sample dictionary of inflected forms automatically produced from a sample dictionary of lemmas and a sample morphological grammars, a syntactic grammar for recognition of one class of named entities – full personal names with their roles or functions, a full compiled dictionary (divided in three files: nouns, verbs, and other). It comprises of 85868 entries: nouns (40886), adjectives (25558), verbs (15366), and other (4058).</p>
Identifier	608
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Computational lexicon
URL	http://korpus.matf.bg.ac.rs/SrpNooJ ili http://www.nooj4nlp.net/pages/resources.html???
Version	v1.0
Last update	2012-08-01

Contacts

Miloš Utvić	
Position	Teaching assistant
Contact	misko@matf.bg.ac.rs
Organization	University of Belgrade Faculty of Philology info@fil.bg.ac.rs

Distribution

Availability	Available – unrestricted use	
IPR holder	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana/
Availability start date	2012-08-01	

Licences

CC-BY		
Restrictions of use	Attribution	
Access medium	Downloadable	
Download location	http://www.nooj4nlp.net/pages/resources.html	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas/
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas/

Metadata

Creation date	2012-07-25	
Metadata creators	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs

	http://poincare.matf.bg.ac.rs/~cvetana
	Ranka Stanković
Position	Assistant professor
Contact	ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en
Source	CESAR
Metadata language ID	en-us
Metadata last date updated	2012-07-25

Validation

Validated	True
Type	Formal
Size	85868 entries
Tool	NooJ;LeXimir
Validator	Miloš Utvić
	Position Teaching Assistant
	Contact misko@matf.bg.ac.rs
	Ranka Stanović
	Position Assitant Professor
	Contact ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en

Usage

Access tool	NooJ;LeXimir
Foreseen use	Human use NLP applications
NLP-specific use	Annotation Morphosyntactic tagging
Actual uses	NLP applications
	NLP-specific use Information extraction Information retrieval
	Reports Cvetana Krstev, Duško Vitas, "Extending the Serbian E-dictionary by using lexical transducers", in Formaliser les langues avec l'ordinateur : De INTEX à Nooj, eds. Svetla Koeva, Denis Maurel, Max Silberstein, pp. 147-168, Presses Universitaires de Franche Comté, Besancon, 2007. Sandra Gucul-Milojević, Vanja Radulović, and Cvetana Krstev. "Usage

	<p>of NooJ Graphs and Annotation for Information Extraction". In Xavier Blanco and Max Silberztein (eds.) Proceedings of the 2007 International Nooj Conference, pp. 103-120, Cambridge Scholars Publishing, 2008. ISBN (13) 978-1-4438-0053-2.</p> <p>Ranka Stanković, Duško Vitas, and Cvetana Krstev. "The Nooj System as Module within an Integrated Language Processing Enironment". In Xavier Blanco and Max Silberztein (eds.) Proceedings of the 2007 International Nooj Conference, pp. 228-248, Cambridge Scolars Publishing, 2008. ISBN (13) 978-1-4438-0053-2.</p> <p>Ranka Stanković, Miloš Utvić, Duško Vitas, Cvetana Krstev, and Ivan Obradović. "On the Compatibility of Lexical Resources for Nooj". In Kristina Vučković, Božo Bekavac and Max Silberztein (eds.) Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the 2011 International Nooj Conference, pp. 96-108, Cambridge Scholars Publishing, 2012. ISBN (13) 978-1-4438-3711-8.</p>	
Usage project	Serbian Language and its Resources: Theory, Description and Applications	
	Project ID	ON 178006
	Funding type	National funds
	Funder	Serbian Ministry of Education and Science
	Country	Serbia
	Start date	2011-01-01
	End date	2015-12-31

Resource creation

Resource creator	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
	Ranka Stanković	
	Position	Assistant professor
	Contact	ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en
	Organization	University of Belgrade

	Faculty of Geology and Mining sladja@rgf.bg.ac.rs
	Miloš Utvić
	Position Teaching assistant
	Contact misko@matf.bg.ac.rs
Funding projects	CEntral And Sout-East Europe An Resources
	Project short name CESAR
	Project ID 271022
	Funding type EU funds
	Start date 2011-02-01
	End date 2013-01-31
Creation start date	2011-04-01

Resource documentation

Samples location	http://www.nooj4nlp.net/pages/resources.html
Tool documentation type	Online

Lexical conceptual resource

Lexical conceptual resource type	Computational lexicon	
Lexical conceptual resource encoding	Encoding level	Morphology
	Linguistic information	Lemma Inflection Part of speech

Texts

Media type	text		
Linguality type	Monolingual		
Languages	Serbian		
	Language ID	srp	
	Language script	Latin/Cyrillic	
	Language variety	Language variety type	Dialect
		Language variety name	Ekavian

		Size	85868 entries
Modality	Modality type	Written language	
Size	85868 entries		
Character encoding	UTF-8		

6.9. Serbian Morphological Dictionary (Multext-East)

General Information

Short name	SrpMD
Description	Morphological electronic dictionary of Serbian (Ekavian pronunciation) (SrpMD) released in the scope of the EU-funded CESAR project is a version of morphological dictionary of Serbian used in the Nooj corpus processing system and constituting the part of the Serbian Nooj Module (see section 6.8). This version is compliant to MULTEXT-East morphosyntactic specification for Serbian (http://nl.ijs.si/ME/V4/msd/html/msd-sr.html) (with one small deviation from it – see section 6.10). It comprises of 3,630,613 entries for 85,721 lemmas covering 11 PoS: nouns (646,867/40,425), adjectives (2,315,640/25,826), verbs (654,159/15,359), adverbs (3233), numerals(4,794/175), conjunctions (83), interjections (218), prepositions (169), pronouns (5,321/104), particles (103), abbreviations (26).
Identifier	609
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Computational lexicon
URL	http://www.korpus.matf.bg.ac.rs/SrpMD/
Version	v1.0
Last update	2012-08-01

Contacts

Cvetana Krstev	
Position	Professor
Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Organization	University of Belgrade Faculty of Philology info@fil.bg.ac.rs

Distribution

Availability	Available – restricted use	
IPR holder	Cvetana Krstev	
	Position	Professor

	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana/
Availability start date	2012-08-01	

Licences

MS-NC-NoReD		
Restrictions of use	Academic – non-commercial use No redistribution	
Access medium	Downloadable	
Download location	http://www.korpus.matf.bg.ac.rs/SrpMD/	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas/
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas/

Metadata

Creation date	2012-07-19	
Metadata creators	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana/
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2012-07-19	

Validation

Validated	True
Type	Formal
Size	3,630,613 entries
Tool	Unitex;Nooj

Validator	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana

Usage

Foreseen use	NLP applications NLP applications	
NLP-specific use	Lemmatization Morphological analysis Morphosyntactic tagging	
Actual uses	NLP applications	
	NLP-specific use	Lemmatization Morphological analysis Morphosyntactic tagging
	Reports	Cvetana Krstev, Processing of Serbian – Automata, Texts and Electronic dictionaries Faculty of Philology, University of Belgrade, Belgrade, 2008.
	Usage project	Serbian Language and its Resources: Theory, Description and Applications
		Project ID ON 178006
		Funding type National funds
		Funder Serbian Ministry of Education and Science
		Country Serbia
		Start date 2011-01-01
		End date 2015-12-31

Resource creation

Resource creator	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Creation start date	1998-01-01	

Resource documentation

Tool documentation type	Online
--------------------------------	--------

Lexical conceptual resource

Lexical conceptual resource type	Computational lexicon	
Lexical conceptual resource encoding	Encoding level	Morphology
	Linguistic information	Lemma Inflection Part of speech
	Conformance to standards best practices	MULTEXT

Texts

Media type	text		
Linguality type	Monolingual		
Languages	Serbian		
	Language ID	srp	
	Language script	Latin	
	Language variety	Language variety type	Dialect
		Language variety name	Ekavian
		Size	3,630,613 entries
Modality	Modality type	Written language	
Size	3,630,613 entries		
Character encoding	UTF-8		

6.10. Morphosyntactically tagged Serbian version of Jules Verne's novel "Around the world in 80 days"

General Information

Short name	Verne80daysMSD
Description	The Serbian version of Jules Verne's novel "Around the world in 80 days" has been automatically tagged and manually disambiguated. Serbian morphological e-dictionaries were used for tagging. The set of morphosyntactic tags used by Serbian e-dictionary were automatically translated to tags conformant to MULTEXT-East morphosyntactic specification for Serbian (http://nl.ijs.si/ME/V4/msd/html/msd-sr.html). There is only a small deviation from this specification concerning the small numbers 2, 3, and 4 (tag 'c' for grammatical number). The final file is conforming to TEI P4 markup of linguistically

	annotated text. Text structure is also tagged: divisions, paragraphs, and segments (sentences). A short TEI header is provided.
Identifier	610
Resource type	Corpus
URL	http://www.korpus.matf.bg.ac.rs/Verne80daysMSD
Version	v1.0
Last update	2012-07-09

Contacts

Cvetana Krstev	
Position	Associate Professor
Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Organization	University of Belgrade Faculty of Philology info@fil.bg.ac.rs

Distribution

Availability	Available – unrestricted use	
IPR holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Availability start date	2012-07-31	

Licences

MS-NC-NoReD		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://www.korpus.matf.bg.ac.rs/Verne80daysMSD	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas

	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Metadata

Creation date	2012-07-09	
Metadata creators	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2012-07-09	

Validation

Validated	True	
Type	Formal	
Mode	Manual	
Extent	Full	
Validator	Miloš Utvić	
	Position	Teaching Assistant
	Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs

Usage

Access tool	Corpus query processor (CQP)
Resource associated with	IMS Open Corpus Workbench (CWB), http://cwb.sourceforge.net
Foreseen use	NLP applications
NLP-specific use	Morphological analysis

	Morphosyntactic tagging Pos tagging	
Actual uses	NLP applications	
	NLP-specific use	Pos tagging
	Reports	<p>Tomaž Erjavec, Cvetana Krstev, Vladimir Petkevič, Kiril Simov, Marko Tadić, Duško Vitas, “The MULTTEXT-East Morphosyntactic Specifications for Slavic Languages”, in Proceedings of the Workshop on Morphological Processing of Slavic Languages : 10th Conference of the European Chapter, EACL 2003, Budapest, Hungary, April 13th, 2003, eds. Tomaž Erjavec and Duško Vitas, pp. 25-32</p> <p>Cvetana Krstev, Duško Vitas, Tomaž Erjavec, “MULTTEXT-East Resources for Serbian”, in Zbornik 7. mednarodne multikonference "Informacijska družba IS 2004", Jezikovne tehnologije 9-15 Oktober 2004, Ljubljana, Slovenija, eds. Tomaž Erjavec, Jerneja Zganec Gros, Institut "Jožef Stefan", Ljubljana, 2004.</p> <p>Dan Tufiş, Svetla Koeva, Tomaž Erjavec, Maria Gavrilidou, and Cvetana Krstev. "Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages". In Marko Tadić, Mila Dimitrova-Vulchanova and Svetla Koeva (eds.) Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008), pp. 145-152, Dubrovnik, Croatia, September 25-28, 2008. ISBN 978-953-55375-0-2.</p> <p>Cvetana Krstev, Duško Vitas, Aleksandra Trtovac, "Orwell's 1984 – the Case of Serbian Revisited", in Proceedings of 5th Language & Technology Conference, November 25-27, 2011, Poznań, Poland, ed. Zygmunt Vetulani, ISBN 978-83-932640-1-8, pp. 570-574, Fundacja Uniwersytetu im. A. Mickiewicza, Poznań, 2011.</p>

Resource creation

Resource creator	Miloš Utvić	
	Position	Teaching Assistant
	Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs
	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Cvetana Krstev	
	Position	Assistant Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Funding projects	SEE-ERA.NET - Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages	

	Project ID	ICT 10503 RP
	Funding type	EU funds
	Funder	European Comission
	Start date	2007-06-01
	End date	2008-06-01
Creation start date	2007-06-01	

Texts

Media type	text		
Linguality type	Monolingual		
Languages	Serbian		
	Language ID	srp	
	Language script	Latin	
	Size	58676 words	
	Language variety	Language variety type	Dialect
		Language variety name	Ekavian
		Size	58676 words
Modality	Modality type	Written language	
Size	58676 words		
Character encoding	UTF-8		
Annotation	Morphosyntactic annotation – POS tagging		
	Annotation standoff	True	
	Segmentation level	Word	
	Format	text/plain	
	Tagset	http://nl.ijs.si/ME/V4/msd/html/msd-sr.html	
	Conformance to standards best practices	MULTEXT	
	Annotation mode	Mixed	
	Annotation tool	Intex and Serbian morphological e-dictionaries	
	Start date	2007-06-01	

	Size	58676 words	
	Annotators	Miloš Utvić	
		Position	Teaching Assistant
		Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs
		Cvetana Krstev	
		Position	Assistant Professor
		Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Domains	literature		
Creation	Original source	retyping	
	Creation mode	Mixed	

6.11. Bibliša: Aligned Collection Search Tool

General Information

Short name	Bibliša
Description	This tool is a web application for search of digital libraries of articles from bilingual e-journals in the form of TMX documents, as well as for development of new bilingual lexical resources based on this search. It is based on previously developed components for LeXimir (work station for lexical resources) and VebRanka (web query expansion tool) and uses various lexical resources: Wordnets, e-dictionaries and terminological lists. Bibliša can expand search queries both morphologically and semantically, as well as to another language, based on available resources. Presently, it is implemented for the Serbian/English bilingual e-journal Infotheca and it uses Serbian morphological e-dictionaries, Serbian and English wordnets connected via the interlingual index, and a bilingual Dictionary of Librarianship. If the search results reveal a shortcoming in existing bilingual resources, an entry to the new bilingual resource is initiated.
Identifier	611
Resource type	Tool/service
Tool/service type	Other
URL	http://cesar.matf.bg.ac.rs/
Version	v1.0
Last update	2012-03-01

Contacts

Ranka Stanković	
Position	Assistant professor
Contact	ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en

Organization	University of Belgrade Faculty of Mining and Geology webmaster@rgf.bg.ac.rs
---------------------	---

Distribution

Availability	Available – restricted use	
IPR holder	Ranka Stanković	
	Position	Assistant professor
	Contact	ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en
Availability start date	2011-12-01	

Licences

GPL		
Restrictions of use	Academic – non-commercial use	
Access medium	Web executable	
Execution location	http://hlt.rgf.bg.ac.rs/Biblisha	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Metadata

Creation date	2012-07-01	
Metadata creators	Ranka Stanković	
	Position	Assistant professor
	Contact	ranka@rgf.bg.ac.rs

	http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en
Source	CESAR
Metadata language ID	en-us
Metadata last date updated	2012-07-01

Validation

Validated	True
Mode	Manual
Validator	Ranka Stanković
	Position Assistant professor
	Contact ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en

Usage

Actual uses	Human use
	<p>Reports Ranka Stanković, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac, Miloš Utvić, A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, May 23-25, 2012</p> <p>Ranka Stanković, Ivan Obradović, Aleksandra Trtovac, An Approach to Development of Bilingual Lexical Resources, CLoBL 2012: Workshop on Computational Linguistics and Natural Language Processing of Balkan Languages, Novi Sad, Sept 16-20, 2012</p>

Resource creation

Resource creator	Ranka Stanković
	Position Assistant professor
	Contact ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en
Funding projects	Infrastructure for E-Learning in Serbia
	Project ID III 47003
	Funding type National funds
	Funder Serbian Ministry of Education and Science
	Country Serbia
	Start date 2011-01-01
	End date 2015-12-31

Creation start date	2011-06-01
---------------------	------------

Resource documentation

Samples location	http://hlt.rgf.bg.ac.rs/biblisha/
Tool documentation type	None

Tool/service

Tool/service type	Other			
Tool/service subtype	Aligned Collection Search Tool			
Language dependent	False			
Input	Media type	text		
Output	Media type	text		
Operating system	Windows			
Required software	NXDB MarkLogic 5			
Required hardware	None			
Required LR	none			
Tool/service evaluation	Evaluated	True		
	Level	Diagnostic		
	Evaluators	Ranka Stanković		
		Position	Assistant professor	
		Contact	ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en	

6.12. Corpus of Contemporary Serbian Newspapers and Magazines

General Information

Short name	SrpNovKor
Description	The Corpus of contemporary Serbian Newspapers and Magazines, SrpNovKor, consists of about 3,3 milion articles. It is not annotated. SrpNovKor texts consist of articles published in period from 2004 to 2012 in following newspapers and magazines: "24 sata", "AG magazin", "Akcija", "Akter", "Alo!", "Ambijenti", "Ana", "Andjela", "Apoteka", "Arena 92", "AS", "Auto", "Auto Bild", "Auto start", "Balkan", "Balkan Ekspres", "Banatske vesti", "Bankar", "Basket", "Bazar", "Bečejski dani", "Bečejski mozaik", "Bečejski dani", "Best", "Best Home", "Best Shop", "Best Shop Kids", "Bilje & zdravlje", "Bilten Regionalne Privredne Komore Užice", "Biznis", "Biznis & finansije", "Biznis i Finansije", "Biznis magazin", "Blic", "Blic Kuhinja", "Blic News", "Blic TV dodatak", "Blic Žena", "Borba", "Borske novine", "Borski Problem", "Brand", "Brand Mania", "Bravacasa", "Bravo", "Čačanske Novine", "Čačanski glas", "Cafe&Bar", "Casaviva", "Centar za modernu politiku", "Cica", "City Magazine", "CKM", "CODE Magazin", "Columbo", "COM",

	<p>"Connect", "CorD", "Cosmopolitan", "Dan", "Danas", "Dani", "Delegacija EU", "Delegacija MON", "DHL biznis info", "Digital", "Digital Foto", "Dijeta", "Dijeta & Lepota", "Divna", "Dnevni Glasnik", "Dnevnik", "Dnevnik - specijalni dodatak", "Dobro jutro", "Doktor u kući", "Dress", "e magazin", "Ekipa", "Ekonometar", "Ekonomist", "Ekonomska politika", "Ekspres", "Elle", "em na kvadrat", "EM portal", "Enterijer", "ESTETSKA hirurgija & kozmetika", "EU market", "Evropa", "Exclusive", "Fame", "FHM", "FHM collection", "Fito Doktor", "Fly & Travel", "Gala", "Gala Style", "Gazeta", "Gazeta - specijalni dodatak", "Gazeta zabava", "Geopolitika", "Gipsware", "Glam Shopping", "Glamur", "Glas - Vrbas", "Glas Javnosti", "Glas komune", "Glas osiguranika", "Glas Podrinja", "Glas Tamnave", "Gloria", "Gloria IN", "Glorija", "GM Business & Lifestyle", "Grad", "Grad Kruševac", "Grad: kulturni vodič kroz Beograd", "Građanski list", "Građanski list - dodatak", "Grazia", "Grom", "Hausbau", "Hej!", "Hello", "Hi-files", "Hrvatska riječ", "Ibarske Novosti", "Ilustrovan politika", "In House", "INDUSTRIJA", "Informer", "Intelligent Life", "Internacional", "Internet ogledalo", "IT market", "JAT Review", "Jelo i piće", "JISA info", "Jolie", "Joy", "Jutarnje ogledalo", "Kikindske", "Kikindske novine", "Knjaževačke Novine", "Kombeg info", "Kontra", "Korak", "Kovinske novine", "Kragujevačke Novine", "Kuća stil", "Kuća Stil+", "Kuhinje & kupatila", "Kulinarske tajne", "Kulska komuna", "Kurir", "Kurir Sport", "Kvart", "kWh", "L'officiel", "Lea", "Lepota i zdravlje", "Lili", "Link", "Lisa", "Local Press", "Lokalna Samouprava", "Lola", "Lozničke Novosti", "M Novine", "Makroekonomske analize i trendovi", "Mama", "Market", "Maxi Magazin", "Maxim", "Mediji o Medijima", "Mega enterijer", "Mens health", "Mikro", "Milenijum", "Mobi", "Mobi Tech", "Mobil MEGA", "Mobilni", "Mobilni magazin", "Moć prirode", "Moda IN", "Modul", "Moj Kutak", "Moj stan", "Moja beba", "Moja Kosa", "Moja lepa bašta", "Moja posla", "Monitor", "Mozzart Sport", "Naftagas promet", "Napred - Valjevo", "Narodne novine - Niš", "Naš Glas", "Naša reč", "Naše novine", "National Geographic - Srbija", "Nautika & Turizam", "Nedeljne novine", "Nedeljni Telegraf", "Nedeljnik", "New Review", "Nezavisna Svetlost", "NIN", "Novi Magazin", "Novi put", "Novokneževačke novine", "NSPM", "O.K.", "Objektiv", "Odbojka Spaja", "Odbrana", "Oficiel", "Ona", "Opozicija", "Optimist", "Palanačke", "Pančevac", "Pančevac pres", "Panorama", "Paparazzo", "PC Magazin", "PC Press", "Pečat", "Pirotske novine", "Playboy", "Pobeda Kruševac", "Pobjeda", "Polimlje", "Politika", "Poljoprivrednik", "Poslovi", "Poslovna Žena", "Poslovne ideje", "Pozorišne novine", "PpresC", "Pravda", "Pravi odgovor", "Pravoslavlje", "Preduzeće", "Pregled", "Press", "Press magazin", "Prestup", "Profil", "Profit", "Progressive magazin", "Prosvetni pregled", "Prosvjetni rad", "Pruga", "Rad Sindikalni Poverenik", "Realno!", "Reč naroda", "ReFoto", "Regionalni dani", "Reporter", "Republika", "Restart", "Revija 024", "Revija 92", "Revija Kolubara", "Revija Uno", "Ribolov", "RIN magazin", "Roditelj & Dete", "Sale & Pepe", "Sat plus", "Savski venac", "Scandal", "Security", "Seljak", "Sensa", "Singidunum Weekly", "Sloboda", "Slobodna reč - Vranje", "Službeni glasnik", "Sofia", "Somborske novine", "Sport", "Sport & Life", "Sport +", "Sportski žurnal", "Srbija", "Sremske novine", "Srpski Nacional", "Stan i kuća", "Standard", "Star", "Stari grad", "Start", "Status", "Stav Naroda", "Story", "Subotičke novine", "Subotički dani", "Sutra", "Svedok", "Svet", "Svet kompjutera", "Svetlost", "T3", "Tabloid", "Taboo", "Takovske novine", "Tenis", "The Best Home", "The Economist", "The Men", "Time Out", "Timok", "Tina", "Top speed", "Travel", "Travel & caffè", "Travel Avantura", "Trudnoća", "TS", "Turbo", "TV Novosti", "U zdravom telu", "Užička nedelja", "Večernje Novosti", "Veliki točkovi", "Vesti Užice", "Vijesti", "Vino", "Vip Trip", "Vita", "Viva", "Vojska", "Vranjske", "Vrele gume", "Vreme", "Vršačka kula", "What HI FI", "Wine Style", "Yachting", "Yellow Cab", "Zakoni", "Zdrav život", "Zdravlje", "Zdravlje i lepota", "Zdravlje i Nauka", "Zdravo dete", "Zdravstveni pregled", "Ženski svet", "Život & Stil", "Zlatarske novosti", "Zrenjanin", "Zrenjaninske novine".</p>
Identifier	612
Resource type	Corpus
URL	http://www.korpus.matf.bg.ac.rs

Version	v1.0
----------------	------

Contacts

Goran Zarić	
Position	Head of Sales Department
Contact	gzaric@arhiv.rs http://poincare.matf.bg.ac.rs/~vitas
Organization	Ebart consulting d.o.o. Media archive office@arhiv.rs

Distribution

Availability	Available – unrestricted use	
IPR holder	Velimir Curgus	
	Position	Director
	Contact	vcurgus@arhiv.rs
	Organization	Ebart consulting d.o.o. Media archive office@arhiv.rs
Availability start date	2012-07-20	

Licences

Proprietary		
Restrictions of use	Commercial use	
Access medium	Downloadable	
Download location	http://www.arhiv.rs/korpus	
Fee	1500 EUR	
Signatories	Velimir Curgus	
	Position	Director
	Contact	vcurgus@arhiv.rs
	Organization	Ebart consulting d.o.o. Media archive office@arhiv.rs
Distribution rights holder	Velimir Curgus	
	Position	Director
	Contact	vcurgus@arhiv.rs
	Organization	Ebart consulting d.o.o.

		Media archive office@arhiv.rs
--	--	---

Metadata

Creation date	2011-11-17	
Metadata creators	Saša Petalinkar	
	Position	IT Developer
	Contact	spetalinkar@arhiv.rs
	Organization	Ebart consulting d.o.o. Media archive office@arhiv.rs
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2012-07-13	

Resource creation

Resource creator	Saša Petalinkar	
	Position	IT Developer
	Contact	spetalinkar@arhiv.rs
	Organization	Ebart consulting d.o.o. Media archive office@arhiv.rs
Creation start date	2012-07-06	

Texts

Media type	text		
Linguality type	Monolingual		
Languages	Serbian		
	Language ID	srp	
	Language script	Latin	
	Size	915772708 words	
	Language variety	Language variety type	Dialect
		Language variety name	Ekavian

		Size	915772708 words
Modality	Modality type	Written language	
Size	915772708 words		
Character encoding	UTF-8		
Creation	Original source	http://www.arhiv.rs	
	Creation mode	Automatic	
	Creation mode details	IMB Lotus Notes	

6.13. Named Entities evaluation corpus for Serbian

General Information

Short name	SrpNEval
Description	Named Entities evaluation corpus for Serbian (SrpNEval) consists of 2000 short news published by Serbian news agencies and daily newspapers in 2005 and 2006. They cover mostly Serbian internal and foreign politics. The size of corpus is: 2,000 short news, 3,343 sentences, 89,425 words, 7,122 named entity tags. Named entities were automatically recognized and manually corrected. Recognized named entities are: persons (full names - 1,648, last names - 152, distinguished - 24), time expressions (dates - 366 and time of day - 33), measurement expressions - 17, money expressions - 146, amount expressions - 322, percent expressions - 53, geopolitical names (countries - 1,390, cities - 1,440, hydronyms - 11, oronyms - 21), organizations - 1,499. The corpus comes in two variants: Latin alphabet and Cyrillic alphabet.
Identifier	613
Resource type	Corpus
URL	http://www.korpus.matf.bg.ac.rs/SrpNEval
Version	v1.0
Last update	2013-01-31

Contacts

Cvetana Krstev	
Position	Professor
Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Organization	University of Belgrade Faculty of Philology info@fil.bg.ac.rs

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

IPR holder	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
	Organization	University of Belgrade Faculty of Philology info@matf.bg.ac.rs
Availability start date	2013-01-31	

Licences

CC-BY-NC		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://www.korpus.matf.bg.ac.rs/SrpNEval	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Metadata

Creation date	2013-01-24	
Metadata creators	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Source	CESAR	
Metadata language ID	en-us	

Metadata last date updated	2013-01-24
-----------------------------------	------------

Validation

Validated	True	
Type	Formal	
Mode	Mixed	
Extent	Full	
Validator	Miloš Utvić	
	Position	Teaching Assistant
	Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs

Usage

Foreseen use	NLP applications	
NLP-specific use	Document classification Named entity recognition Person recognition Temporal expression recognition	
Actual uses	NLP applications	
	NLP-specific use	Information extraction Information retrieval
	Reports	Cvetana Krstev, Jelena Jaćimović and Duško Vitas, "Recognition and normalization of some classes of named entities in Serbian", BCI '12 Proceedings of the Fifth Balkan Conference in Informatics, ACM New York, NY, USA, 2012, pp. 52-57, eds. Mirjana Ivanović and Zoran Budimac, ISBN 978-1-4503-1240-0, DOI 10.1145/2371316.2371327 Cvetana Krstev, Duško Vitas, Ivan Obradović, Miloš Utvić, "E-Dictionaries and Finite-State Automata for the Recognition of Named Entities", in Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing, FSMNLP 2011, Blois, France, July 12-15, 2011. eds. Andreas Maletti and Matthieu Constant, Association for Computational Linguistics, ISBN 978-3-642-14769-2, pp. 48-56, 2011.

Resource creation

Resource creator	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana

	Miloš Utvić	
	Position	Teaching Assistant
	Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs
Funding projects	#47003 - Infrastructure for Electronically Supported Learning in Serbia	
	Project ID	#47003
	Funding type	National funds
	Funder	Serbian Ministry of Education and Science
	Country	Serbia
	Start date	2011-01-01
	End date	2015-12-31
Creation start date	2011-01-01	

Texts

Media type	text		
Linguality type	Monolingual		
Languages	Serbian		
	Language ID	srp	
	Language script	Latin	
	Size	89425 words	
	Language variety	Language variety type	Dialect
		Language variety name	Ekavian
		Size	89425 words
Modality	Modality type	Written language	
Size	89425 words		
Character encoding	UTF-16		
Annotation	Semantic annotation – named entities		
	Annotation standoff	True	
	Segmentation level	Sentence	
	Format	XML	
	Annotation mode	Mixed	

	Annotation tool	http://hlt.rgf.bg.ac.rs/VeBranka/NERanka.aspx
	Start date	2011-01-01
	Size	7122 elements
	Annotators	Cvetana Krstev
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
	Miloš Utvić	
	Position	Teaching Assistant
	Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs
Domains	news	
Creation	Original source	http://www.arhiv.rs
	Creation mode	Mixed

6.14. Serbian NGrams

General Information

Short name	SrpNGrams
Description	Serbian NGrams (SrpNGrams) represent set of N-grams extracted from Serbian Lemmatized and PoS Annotated Corpus (SrpLemKor) for N from 1 to 5. Each unigram is maximum continuous chunk of non-whitespace lower-case characters. The resource contains all unique N-grams preceded by number of occurrences. It also contains n-gram language models (1-5) in the standard ARPA text and binary format, created by IRST Language Modeling Toolkit. SrpKor texts consist of: fiction written by Serbian authors in 20th and 21th century, various scientific texts from various domains (both humanities and sciences), legislative texts and general texts. General texts represent daily news published in newspaper "Politika" 2000-2002 and 2005-2010, texts in journals and magazines 1991-2002 ("Danica", "Ebit", "Ekonomist", "Glasnik", "NIN", "Ilustrovana politika", "Kalibar", "Moje srce", "Mostovi", "Pravoslavlje", "Svet", "Teološki pogledi", "Trn", "Viva", "Republika"), internet portal texts 2011-2012 (Peščanik), TANJUG agency news 1995-96, newspaper feuilletons published in newspapers "Politika" (2001-2003), "Večernje novosti" (2008-2011) and "Danas" (2002-2006).
Identifier	614
Resource type	Corpus
URL	http://korpus.matf.bg.ac.rs/SrpNgrams
Version	v1.0
Last update	2013-01-31

Contacts

Duško Vitas

Position	Professor
Contact	vas@matf.bg.ac.rs http://www.matf.bg.ac.rs/~vas
Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Mirko Spasić	
Position	Teaching assistant
Contact	mirko@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~mirko

Distribution

Availability	Available – restricted use	
IPR holder	Duško Vitas	
	Position	Professor
	Contact	vas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
	Mirko Spasić	
	Position	Teaching assistant
	Contact	mirko@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~mirko
Availability start date	2013-01-23	

Licences

MS-NC-NoReD-ND		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://korpus.matf.bg.ac.rs/SrpNgrams	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vas
	Organization	University of Belgrade Faculty of Mathematics

		matf@matf.bg.ac.rs
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Metadata

Creation date	2013-01-31	
Metadata creators	Mirko Spasić	
	Position	Teaching Assistant
	Contact	mirko@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~mirko
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2013-01-31	

Validation

Validated	True	
Type	Formal	
Validator	Mirko Spasić	
	Position	Teaching assistant
	Contact	mirko@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~mirko

Resource creation

Resource creator	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
	Mirko Spasić	
	Position	Teaching assistant

	Contact	mirko@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~mirko
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	Project ID	271022
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-12-15	

Resource documentation

Tool documentation type	Online
--------------------------------	--------

Corpus text ngram

Media type	textNgram	
Ngram	Base item	Word
	Order	5
Linguality type	Monolingual	
Languages	Serbian	
	Language ID	srp
Modality	Modality type	Written language
Size	402395 unigrams, 2148570 bigrams, 3309528 trigrams, 3664760 4 – grams, 3740032 5 – grams	

6.15. NERanka: Named Entity Recognition and Annotation Tool

General Information

Short name	NERanka
Description	This tool is a web application for named entity (NE) recognition and tagging in Serbian that relies on large-scale lexical resources and a cascade of finite-state transducers. It is based on previously developed components for LeXimir (Tool for lexical resources management and query expansion) and VebRanka (web query expansion tool) and uses various lexical resources. NERanka recognizes several types of named entities, and the user can choose either all NE types or select specific NEs from the Person, Organization, Location, Amount or Time group. Annotation of uploaded text is possible for XML and textual documents. For tagging NERanka uses XML tags which describe the type and subtype of the named entity.
Identifier	615

Resource type	Tool/service
Tool/service type	Other
URL	http://hlt.rgf.bg.ac.rs/VeBranka/NERanka.aspx
Version	v1.0
Last update	2013-01-20

Contacts

Cvetana Krstev	
Position	Professor
Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Organization	University of Belgrade Faculty of Philology info@fil.bg.ac.rs
Ranka Stanković	
Position	Assistant professor
Contact	ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en
Organization	University of Belgrade Faculty of Mining and Geology webmaster@rgf.bg.ac.rs

Distribution

Availability	Available – restricted use	
IPR holder	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
	Ranka Stanković	
	Position	Assistant professor
	Contact	ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en
Availability start date	2012-12-21	

Licences

GPL	
Restrictions of use	Academic – non-commercial use
Access medium	Web executable

Execution location	http://hlt.rgf.bg.ac.rs/VeBranka/NERanka.aspx	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Metadata

Creation date	2013-01-20	
Metadata creators	Ranka Stanković	
	Position	Assistant professor
	Contact	ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2013-01-20	

Validation

Validated	True	
Mode	Manual	
Validator	Ranka Stanković	
	Position	Assistant professor
	Contact	ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en

Usage

Actual uses	Human use
--------------------	-----------

	<p>Reports</p> <p>Cvetana Krstev, Jelena Jaćimović and Duško Vitas, “Recognition and normalization of some classes of named entities in Serbian”, BCI '12 Proceedings of the Fifth Balkan Conference in Informatics, ACM New York, NY, USA, 2012, pp. 52-57, eds. Mirjana Ivanović and Zoran Budimac, ISBN 978-1-4503-1240-0, DOI 10.1145/2371316.2371327</p> <p>Cvetana Krstev, Duško Vitas, Ivan Obradović, Miloš Utvić, “E-Dictionaries and Finite-State Automata for the Recognition of Named Entities”, in Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing, FSMNLP 2011, Blois, France, July 12-15, 2011. eds. Andreas Maletti and Matthieu Constant, Association for Computational Linguistics, ISBN 978-3-642-14769-2, pp. 48-56, 2011.</p>
--	---

Resource creation

Resource creator	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
	Ranka Stanković	
	Position	Assistant professor
	Contact	ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en
Funding projects	Infrastructure for E-Learning in Serbia	
	Project ID	III 47003
	Funding type	National funds
	Funder	Serbian Ministry of Education and Science
	Country	Serbia
	Start date	2011-01-01
	End date	2015-12-31
	CEntral And Sout-East EuropeAn Resources	
	Project short name	CESAR
	Project ID	271022
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-31
	Creation start date	2012-06-01

Resource documentation

--	--

Samples location	http://hlt.rgf.bg.ac.rs/VeBranka/NERanka.aspx
Tool documentation type	None

Tool/service

Tool/service type	Other			
Tool/service subtype	Annotation Tool			
Language dependent	False			
Input	Media type	text		
Output	Media type	text		
Operating system	Windows			
Required software	Unitex 3.0			
Required hardware	None			
Required LR	none			
Tool/service evaluation	Evaluated	True		
	Level	Diagnostic		
	Evaluators	Ranka Stanković		
		Position	Assistant professor	
		Contact	ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en	

6.16. Serbian Spell Checker

General Information

Short name	SrpSpell
Description	Serbian Spell Checker (SrpSpell) contains three dictionaries (Cyrillic, Latin and combined) and two word lists (Cyrillic and Latin). Combined dictionary is default. Words are encoded in UTF-8 code page, normalized to special 8bit code page l-sr. Description of this code page can be found in file misc/l-sr.txt. GNU aspell files l-sr.cset and l-sr.cmap are included in this package. Support for accented vowels is built in, but in current release dictionary does not have any words accented vowels.
Identifier	616
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Word list
URL	http://korpus.matf.bg.ac.rs/SrpSpell
Version	v0.02
Last update	2005-09-11

Contacts

Duško Vitas	
Position	Professor
Contact	vitas@matf.bg.ac.rs http://www.matf.bg.ac.rs/~vitas
Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Distribution

Availability	Available – restricted use	
IPR holder	Goran Rakić	
	Contact	grakic@devbase.net http://blog.goranrakic.com/
Availability start date	2013-02-01	

Licences

LGPL		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://korpus.matf.bg.ac.rs/SrpSpell	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas/
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas/

Metadata

Creation date	2013-01-22	
Metadata creators	Miloš Utvić	
	Position	Teaching Assistant
	Contact	misko@matf.bg.ac.rs

Source	CESAR
Metadata language ID	en-us
Metadata last date updated	2012-07-27

Validation

Validated	True	
Type	Formal	
Size	343242 words	
Validator	Miloš Utvić	
	Position	Teaching Assistant
	Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs

Usage

Access tool	GNU Aspell, http://aspell.net/	
Foreseen use	Human use NLP applications	
NLP-specific use	Lexicon acquisition from corpora Lexicon enhancement	
Actual uses	NLP applications	
	NLP-specific use	Spell checking
	Usage project	Serbian Language and its Resources: Theory, Description and Applications
		Project ID ON 178006
		Funding type National funds
		Funder Serbian Ministry of Education and Science
		Country Serbia
		Start date 2011-01-01
		End date 2015-12-31

Resource creation

Resource creator	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs

	http://poincare.matf.bg.ac.rs/~vitas
	Organization University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
	Goran Rakić
	Contact grakic@devbase.net http://blog.goranrakic.com/
Funding projects	Localization of OpenOffice for Serbian
	Funding type National funds
	Funder Serbian Ministry of Telecommunication
	Country Serbia
	Start date 2008-01-01
	End date 2009-12-31
Creation start date	2008-01-01

Resource documentation

Tool documentation type	Online
--------------------------------	--------

Lexical conceptual resource

Lexical conceptual resource type	Word list	
Lexical conceptual resource encoding	Encoding level	Morphology
	Linguistic information	Other

Texts

Media type	text		
Linguality type	Monolingual		
Languages	Serbian		
	Language ID	srp	
	Language script	Latin	
	Language variety	Language variety type	Dialect
		Language variety name	Ekavian
		Size	343242 words

Modality	Modality type	Written language
Size	343242 words	
Character encoding	UTF-8	

6.17. Emotions Annotation Tool

General Information

Short name	eEmotion
Description	This tool is a web application for ontological based emotions recognition and tagging of Serbian texts. The application uses RDFS which are created by using nine discrete emotion psychological theories. Also, it uses associative dictionary of Serbian with about 11 thousands words and Serbian morphological electronic dictionary which contains approximately 4.4 million different inflectional forms of simple words. The application offers a representation of summary results in a graphical form. Annotation of an uploaded text is possible for XML and textual documents as well as a text from Web.
Identifier	617
Resource type	Tool/service
Tool/service type	Other
URL	http://cvetana.mmiljana.com/
Version	v1.1
Last update	2012-05-15

Contacts

Cvetana Krstev	
Position	Professor
Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Organization	University of Belgrade Faculty of Philology info@fil.bg.ac.rs
Miljana Mladenović	
Position	PhD student
Contact	ml.miljana@gmail.com http://cvetana.mmiljana.com
Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Distribution

--	--

Availability	Available – restricted use	
IPR holder	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
	Miljana Mladenović	
	Position	PhD student
	Contact	ml.miljana@gmail.com http://cvetana.mmiljana.com
Availability start date	2012-12-21	

Licences

GPL		
Restrictions of use	Academic – non-commercial use	
Access medium	Web executable	
Execution location	http://cvetana.mmiljana.com	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Metadata

Creation date	2013-01-31	
Metadata creators	Miljana Mladenović	
	Position	PhD student
	Contact	ml.miljana@gmail.com http://cvetana.mmiljana.com

Source	CESAR
Metadata language ID	en-us
Metadata last date updated	2013-01-31

Validation

Validated	True		
Mode	Manual		
Validator	Miljana Mladenović		
	Position	PhD student	
	Contact	ml.miljana@gmail.com http://cvetana.mmiljana.com	

Usage

Actual uses	Human use
--------------------	-----------

Resource creation

Resource creator	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
	Miljana Mladenović	
	Position	PhD student
	Contact	ml.miljana@gmail.com http://cvetana.mmiljana.com
Creation start date	2011-12-20	

Resource documentation

Samples location	http://cvetana.mmiljana.com
Tool documentation type	None

Tool/service

Tool/service type	Other
Tool/service subtype	Emotion Annotation Tool
Language dependent	False

Input	Media type	text
Output	Media type	text
Operating system	Windows	
Required hardware	None	
Required LR	none	
Tool/service evaluation	Evaluated	True
	Level	Diagnostic
	Evaluators	Miljana Mladenović
		Position PhD student
		Contact ml.miljana@gmail.com http://cvetana.mmiljana.com

6.18. NERosetta

General Information

Short name	NERosetta
Description	NERosetta is a multiuser web application that aims to facilitate retrieval and comparison of named entities in a single or parallel texts. The main named entity categorization is realized according to the Quaero annotation recommendation and provides a user with approximately 50 different search options. Registered users have an extra possibility to share annotated resources (in XML format) and annotation schemas for the set of languages as well as to manage their own resources and schemas. The initial version supports four annotation schemas (Stanford NER 3 and Stanford NER 7 for English, Krstev&Vitas for Serbian and Maurel for French) and three annotated parallel versions of Jules Verne's Around The World in Eighty Days (English-Serbian, French-Serbian and French-English).
Identifier	618
Resource type	Tool/service
Tool/service type	Other
URL	http://arhimed.matf.bg.ac.rs/~andjelka/paralel-extended
Version	v1.0
Last update	2013-01-22

Contacts

Cvetana Krstev	
Position	Professor
Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Organization	University of Belgrade

	Faculty of Philology info@fil.bg.ac.rs
Andelka Zečević	
Position	Teaching assistant
Contact	andjelkaz@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~andjelkaz
Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Distribution

Availability	Available – restricted use	
IPR holder	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
	Andelka Zečević	
	Position	Teaching assistant
	Contact	andjelkaz@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~andjelkaz
Availability start date	2013-01-31	

Licences

GPL		
Restrictions of use	Academic – non-commercial use	
Access medium	Web executable	
Execution location	http://arhimed.matf.bg.ac.rs/~andjelka/paralel-extended	
Fee	no price	
Signatories	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs
Distribution rights holder	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs

		http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Metadata

Creation date	2013-01-22	
Metadata creators	Anđelka Zečević	
	Position	Teaching assistant
	Contact	andjelkaz@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~andjelkaz
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2013-01-22	

Validation

Validated	True	
Mode	Manual	
Validator	Anđelka Zečević	
	Position	Teaching assistant
	Contact	andjelkaz@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~andjelkaz

Resource creation

Resource creator	Anđelka Zečević	
	Position	Teaching assistant
	Contact	andjelkaz@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~andjelkaz
	Krstev Cvetana	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Funding projects	CeNtral And Sout-East EuropeAn Resources	
	Project short name	CESAR
	Project ID	271022
	Funding type	EU funds

	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2012-05-11	

Resource documentation

Samples location	http://arhived.matf.bg.ac.rs/~andjelka/paralel_extended
Tool documentation type	None

Tool/service

Tool/service type	Other			
Tool/service subtype	Aligned Texts Search			
Language dependent	False			
Input	Media type	text		
Output	Media type	text		
Operating system	Linux			
Required software	LAMP platform			
Required hardware	None			
Required LR	none			
Tool/service evaluation	Evaluated	True		
	Level	Diagnostic		
	Evaluators	Anđelka Zečević		
		Position	Teaching assistant	
		Contact	andjelkaz@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~andjelkaz	

6.19. Rhetorical Figures for Serbian

General Information

Short name	SerRetFig
Description	Rhetorical Figures is a database for Serbian that consists of 98 rhetorical figures related to rehetorical figures for English located at http://rhetfig.appspot.com/list . It is downloadable in xml format. The RhetFig tool is created for maintaining the database, adding examples in it and sorting by: rhetorical types, linguistic types or linguistic operations.
Identifier	619
Resource type	Lexical conceptual resource

Lexical conceptual resource type	Other
URL	http://resursi.mmiljana.com/Data/retFigSrp.xml
Version	v1.0
Last update	2012-12-01

Contacts

Miljana Mladenović	
Position	PhD student
Contact	ml.miljana@gmail.com http://resursi.mmiljana.com
Organization	University of Belgrade Faculty of Mathematics matf@matf.bg.ac.rs

Distribution

Availability	Available – restricted use	
IPR holder	Miljana Mladenović	
	Position	PhD student
	Contact	ml.miljana@gmail.com http://resursi.mmiljana.com
Availability start date	2012-12-01	

Licences

MS-NC-NoReD		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://resursi.mmiljana.com/	
Fee	no price	
Signatories	Miljana Mladenović	
	Position	PhD student
	Contact	ml.miljana@gmail.com http://resursi.mmiljana.com
Distribution rights holder	Miljana Mladenović	
	Position	PhD student
	Contact	ml.miljana@gmail.com http://resursi.mmiljana.com

Metadata

Creation date	2012-12-01	
Metadata creators	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2012-12-01	

Validation

Validated	True	
Type	Formal	
Size	98 idiomatic expressions	
Tool	http://www.xmlvalidation.com/	
Validator	Miljana Mladenović	
	Position	PhD student
	Contact	ml.miljana@gmail.com http://resursi.mmiljana.com

Usage

Access tool	http://resursi.mmiljana.com	
Foreseen use	Human use	
	NLP applications	
NLP-specific use	Discourse analysis	
	Sentiment analysis	
Actual uses	NLP applications	
	NLP-specific use	Information extraction Information retrieval

Resource creation

Resource creator	Miljana Mladenović	
	Position	PhD student
	Contact	ml.miljana@gmail.com http://resursi.mmiljana.com

Creation start date	2012-08-01
----------------------------	------------

Lexical conceptual resource

Lexical conceptual resource type	Other	
Lexical conceptual resource encoding	Encoding level	Semantics
	Linguistic information	Definition/gloss Usage – examples Usage – notes

Texts

Media type	text		
Linguality type	Monolingual		
Languages	Serbian		
	Language ID	srp	
	Language script	Latin	
	Language variety	Language variety type	Dialect
		Language variety name	Ekavian
		Size	98 idiomatic expressions
Modality	Modality type	Written language	
Size	98 idiomatic expressions		
Character encoding	UTF-8		

7. PUPIN resources

7.1. MONO version of NooJ

General Information

Short name	MONONooJ
Description	NooJ is a linguistic development environment that allows linguists to formalize several levels of linguistic phenomena: typography and spelling; lexicons of simple words, multiword units and discontinuous expressions; inflectional, derivational and productive morphology; local and structural syntax, transformational and semantic analysis and generation. For each of these levels NooJ provides linguists with one formal framework specifically designed to facilitate the description of each phenomenon, as well as parsing/development/debugging tools designed to be as computationally efficient as possible, from Finite-State machines to

	Turing machines. This approach distinguishes NooJ from other computational linguistic frameworks which provide a unique formalism based on a compromise between power and efficiency. As a corpus processing tool, NooJ allows all researchers and professional to extract information from general or technical corpora by applying sophisticated queries based on concepts rather than word forms and build indices, add semantic annotations, perform statistical analyses, etc. MONO version of NooJ is operative on all platforms that support MONO.
Identifier	701
Resource type	Tool/service
Tool/service type	Nlp development environment
URL	http://www.nooj4nlp.net/pages/download.html
Version	v1.0
Last update	2013-01-23

Contacts

Max Silberstein	
Position	Professor
Contact	elliadd@univ-fcomte.fr http://elliadd.univ-fcomte.fr/
Organization	Université de Franche-Comté FELLIAD elliadd@univ-fcomte.fr

Distribution

Availability	Available – restricted use	
IPR holder	Max Silberstein	
	Position	Professor
	Contact	Max.Silberstein@univ-fcomte.fr http://www.nooj4nlp.net/pages/author.html
	Organization	Université de Franche-Comté FELLIAD elliadd@univ-fcomte.fr
Availability start date	2013-01-23	

Licences

MS-NC-NoReD-ND	
Restrictions of use	Academic – non-commercial use No derivatives No redistribution

Access medium	Downloadable	
Download location	http://www.nooj4nlp.net/pages/download.html	
Fee	no price	
Signatories	Max Silberztein	
	Position	Professor
	Contact	Max.Silberztein@univ-fcomte.fr http://www.nooj4nlp.net/pages/author.html
	Organization	Université de Franche-Comté ELLIAD elliadd@univ-fcomte.fr
Distribution rights holder	Max Silberztein	
	Position	Professor
	Contact	Max.Silberztein@univ-fcomte.fr http://www.nooj4nlp.net/pages/author.html
	Organization	Université de Franche-Comté FELLIAD elliadd@univ-fcomte.fr

Metadata

Creation date	2013-01-23	
Metadata creators	Mladen Stanojević	
	Position	Scientific Advisor
	Contact	Mladen.Stanojevic@pupin.rs http://www.pupin.rs/en/imp-organization/fraunhofer-pupin-jpo/jpo-team/#mladenstojanovic
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2013-01-23	

Validation

Validated	True	
Mode	Manual	
Validator	Max Silberztein	
	Position	Professor
	Contact	Max.Silberztein@univ-fcomte.fr http://www.nooj4nlp.net/pages/author.html

Resource creation

Resource creator	Mladen Stanojević	
	Position	Scientific Advisor
	Contact	Mladen.Stanojevic@pupin.rs http://www.pupin.rs/en/imp-organization/fraunhofer-pupin-jpo/jpo-team/#mladenstojanovic
Funding projects	CEntral and South-East European Resources	
	Project ID	271022
	Funding type	EU funds
	Funder	EC Competitiveness and Innovation Framework Programme
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2011-02-01	

Resource documentation

Samples location	http://www.nooj4nlp.net/pages/references.html
Tool documentation type	Manual

Tool/service

Tool/service type	Nlp development environment			
Language dependent	False			
Input	Media type	text		
Output	Media type	text		
Operating system	OS-independent			
Required software	http://www.mono-project.com/Main_Page			
Required hardware	None			
Required LR s	none			
Tool/service evaluation	Evaluated	True		
	Level	Usage		
	Evaluators	Max Silberztein		
		Position	Professor	
		Contact	Max.Silberztein@univ-fcomte.fr http://www.nooj4nlp.net/pages/author.html	

7.2. Java version of NooJ

General Information

Short name	JavaNooJ
Description	NooJ is a linguistic development environment that allows linguists to formalize several levels of linguistic phenomena: typography and spelling; lexicons of simple words, multiword units and discontinuous expressions; inflectional, derivational and productive morphology; local and structural syntax, transformational and semantic analysis and generation. For each of these levels NooJ provides linguists with one formal framework specifically designed to facilitate the description of each phenomenon, as well as parsing/development/debugging tools designed to be as computationally efficient as possible, from Finite-State machines to Turing machines. This approach distinguishes NooJ from other computational linguistic frameworks which provide a unique formalism based on a compromise between power and efficiency. As a corpus processing tool, NooJ allows all researchers and professional to extract information from general or technical corpora by applying sophisticated queries based on concepts rather than word forms and build indices, add semantic annotations, perform statistical analyses, etc. Java version of NooJ is an open source software and working on all operating systems.
Identifier	702
Resource type	Tool/service
Tool/service type	Nlp development environment
URL	http://www.nooj4nlp.net/pages/download.html
Version	v1.0
Last update	2013-01-23

Contacts

Max Silberztein	
Position	Professor
Contact	elliadd@univ-fcomte.fr http://elliadd.univ-fcomte.fr/
Organization	Université de Franche-Comté FELLIAD elliadd@univ-fcomte.fr

Distribution

Availability	Available – restricted use	
IPR holder	Max Silberztein	
	Position	Professor
	Contact	Max.Silberztein@univ-fcomte.fr http://www.nooj4nlp.net/pages/author.html
	Organization	Université de Franche-Comté FELLIAD elliadd@univ-fcomte.fr
Availability start date	2013-03-01	

Licences

GPL		
Restrictions of use	Academic – non-commercial use	
Access medium	Downloadable	
Download location	http://www.nooj4nlp.net/pages/download.html	
Fee	no price	
Signatories	Max Silberztein	
	Position	Professor
	Contact	Max.Silberztein@univ-fcomte.fr http://www.nooj4nlp.net/pages/author.html
	Organization	Université de Franche-Comté ELLIAD elliadd@univ-fcomte.fr
Distribution rights holder	Max Silberztein	
	Position	Professor
	Contact	Max.Silberztein@univ-fcomte.fr http://www.nooj4nlp.net/pages/author.html
	Organization	Université de Franche-Comté FELLIAD elliadd@univ-fcomte.fr

Metadata

Creation date	2013-01-23	
Metadata creators	Mladen Stanojević	
	Position	Scientific Advisor
	Contact	Mladen.Stanojevic@pupin.rs http://www.pupin.rs/en/imp-organization/fraunhofer-pupin-jpo/jpo-team/#mladenstojanovic
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2013-01-23	

Validation

Validated	True	
Mode	Manual	
Validator	Max Silberztein	

	Position	Professor
	Contact	Max.Silberstein@univ-fcomte.fr http://www.nooj4nlp.net/pages/author.html

Resource creation

Resource creator	Mladen Stanojević	
	Position	Scientific Advisor
	Contact	Mladen.Stanojevic@pupin.rs http://www.pupin.rs/en/imp-organization/fraunhofer-pupin-jpo/jpo-team/#mladenstojanovic
Funding projects	CEntral and South-East European Resources	
	Project ID	271022
	Funding type	EU funds
	Funder	EC Competitiveness and Innovation Framework Programme
	Start date	2011-02-01
	End date	2013-01-31
Creation start date	2011-12-01	

Resource documentation

Samples location	http://www.nooj4nlp.net/pages/references.html
Tool documentation type	Manual

Tool/service

Tool/service type	Nlp development environment		
Language dependent	False		
Input	Media type	text	
Output	Media type	text	
Operating system	OS-independent		
Required software	http://www.java.com		
Required hardware	None		
Required LRs	none		
Tool/service evaluation	Evaluated	True	
	Level	Usage	
	Evaluators	Max Silberztein	
		Position	Professor

		Contact	Max.Silberstein@univ-fcomte.fr http://www.nooj4nlp.net/pages/author.html
--	--	----------------	---

8. IBL resources

8.1. Bulgarian National Corpus

General Information

Short name	BulNC
Description	<p>The Bulgarian National Corpus (BulNC) is a large representative publicly available corpus. It is designed as a uniform framework for texts of different modality (written and spoken), period, and number of languages (monolingual and parallel). Its core incorporates several electronic corpora, developed in the period 2001-2009 but has been substantially expanded in the following years. The corpus reflects the state of the Bulgarian language (mainly in its written form) from 1945 until the present. The enlargement of the BulNC has involved not only the amassing of Bulgarian texts, but also the compilation of parallel corpora with Bulgarian as a pivot language. The texts in other languages obligatory have a Bulgarian counterpart in the Bulgarian part of the corpus. Currently, the corpus core consists of over 1.2 billion words and about 240,000 texts. So far 47 foreign languages have been included totalling about 4.2 billion words. Thus, the overall size of the corpus exceeds 5.4 billion words. All texts are supplied with extensive metadata description compliant with the established standards. The corpus is supplied with three levels of annotation:</p> <ul style="list-style-type: none"> • A detailed metadata description: each text is supplied with editorial (author's name, text title, source, etc.) and classificatory metadata (general category, domain, genre). • Monolingual annotation: tokenisation, sentence splitting, POS tagging, lemmatisation, word sense annotation. • Multilingual annotation: alignment at different levels, currently sentence and clause level. <p>The tagset used in the annotation of the BulNC is available as the Bulgarian tagset. The Bulgarian part and the Bulgarian-English parallel corpus are tokenised, sentence-split, POS tagged and lemmatised; the Bulgarian part is also word sense annotated. For the time being, the corpora for the other languages are tokenised, sentence-split and aligned. A special corpus search system allows complex queries to be performed. A set of tools was developed for extracting the metadata and compiling the corpus description from the markup formats. The metadata are as detailed as possible in order to ensure easy text classification, corpus evaluation, derivation of subcorpora based on a set of criteria (e.g. publishing year, domain), and others. The Bulgarian National Corpus Collocation service (http://dcl.bas.bg/collocations/?cmd=collocations&word=%D0%BD%D0%B5%D1%82) gives access to the Bulgarian National Corpus. The service employs the free-of-charge NoSketchEngine, a system for corpora processing that combines Manatee and Bonito. The Collocation service is a RESTful webservice, supporting complicated queries through http. Example: http://dcl.bas.bg/collocations/?cmd=collocations&word=heruser:bulncpass:bulnc The query returns the collocations of a given word in the NoSketchEngine format. The system also supports additional arguments, namely all that are accepted by NoSketchEngine, provided with default values.</p>
Identifier	801
Resource type	Corpus
URL	http://ibl.bas.bg/en/BGNC_en.htm http://search.dcl.bas.bg/

Version	5.0
Last update	2013-01-20

Contacts

Angel Genov	
Position	Affiliated researcher
Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia dcltools@dcl.bas.bg http://dcl.bas.bg/PersonalPages/angel/
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of computational linguistics, Department of Bulgarian lexicology and lexicography
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg
Availability start date	2008-02-01	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	http://search.dcl.bas.bg
Restrictions of use	Academic – non-commercial use
Access medium	Web executable
Execution location	http://dcl.bas.bg/collocations/?cmd=collocations&word=HET

Metadata

Creation date	2011-11-20
Metadata last date	2013-02-01

updated	
----------------	--

Validation

Validated	True
------------------	------

Usage

Access tool	http://dcl.bas.bg/BulNC-registration/?lang=BG
Foreseen use	Human use NLP applications
Actual uses	Human use NLP applications

Resource creation

Resource creator	Institute for Bulgarian Language	
	Department name	Department of computational linguistics, Department of Bulgarian lexicology and lexicography
	Contact	bulnc@dcl.bas.bg http://ibl.bas.bg/en/BGNC_en.htm
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds Own funds
	Start date	2011-02-01
	End date	2013-01-30
	Bulgarian National Corpus project	
	Project ID	BulNC
	URL	http://ibl.bas.bg/en/BGNC_en.htm
	Funding type	National funds
	Country	Bulgaria
	Start date	2009-12-17
	End date	2013-06-17

Resource documentation

Reports	Koeva, Svetla, Ivelina Stoyanova, Svetlozara Leseva, Tsvetana Dimitrova, Rositsa Dekova, Ekaterina Tarpomanova. The Bulgarian National Corpus: Theory and practice in corpus design. – Journal of Language Modelling, 2012, 1 (1), pp. 65-110. ISSN: 2299-
----------------	--

	8470. Koeva, Svetla, Diana Blagoeva, Sia Kolkovska. Levels of annotation in the Bulgarian National Corpus. – Prace Filologiczne, 2012, LXIII, pp. 147-153. ISSN: 0138-0567. Blagoeva, Diana, Sia Kolkovska, Nadezhda Kostova, Cvetelina Georgieva. The Bulgarian National Corpus and its application in Bulgarian academic lexicography. – Prace Filologiczne, 2012, LXIII, pp. 37-49. ISSN: 0138-0567. Koeva, Svetla, Angel Genov. Bulgarian language processing chain. In Proceedings of Integration of Multilingual Resources and Tools in Web Applications. Proceedings of a Workshop in conjunction with GSCL 2011, University of Hamburg, 2011.
Tool documentation type	Help functions Manual Online

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
Modality	Modality type	Written language
Size	1,202,209,147 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph Sentence Word
	Morphosyntactic annotation – POS tagging	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Semantic annotation	
	Segmentation level	Word
	Semantic annotation – word senses	
	Segmentation level	Word

8.2. The Bulgarian National Corpus Collocation service

General Information

Description	<p>The Bulgarian National Corpus Collocation service (http://dcl.bas.bg/collocations/?cmd=collocations&word=%D0%BD%D0%B5%D1%82) gives access to the Bulgarian National Corpus. The service employs the free-of-charge NoSketchEngine, a system for corpora processing that combines Manatee and Bonito. The Collocation service is a RESTful webservice, supporting complicated queries through http.</p> <p>user: bulnc pass: bulnc</p> <p>The query returns the collocations of a given word in the NoSketchEngine format.</p> <p>The system also supports additional arguments, namely all that are accepted by NoSketchEngine, provided with default values.</p> <p>The Collocation service is one of the 4 different means of access to the Bulgarian National Corpus.</p>
--------------------	--

8.3. Bulgarian Part-of-Speech Corpus

General Information

Short name	BulPosCor
Description	<p>The Bulgarian Part-of-Speech Corpus (BulPosCor) is derived from the Brown Corpus of Bulgarian, automatically annotated respectively with PoS tags and manually disambiguated. The corpus for annotation was built by selecting portions of 150+ words from each sample from the Brown Corpus of Bulgarian. The automatic grammatical annotation of the corpus employed the Bulgarian Grammar Dictionary containing about 85 000 words and over 1.5 million word forms specified with grammatical characteristics. Disambiguation was performed by human experts that assigned the correct PoS tags out of two or more possible for an ambiguous token. A number of annotation principles had been outlined in order to provide a uniform approach to the annotation. As a result a PoS disambiguated corpus was obtained consisting of 217 210 tokens, including 172 482 single words, 42 058 punctuation marks and 2 670 numbers. The chief intended application of the Bulgarian Tagged Corpora is to serve as a test and/or training dataset for PoS disambiguation. The Tagged Corpus enables efficient online search of language patterns and forms as well.</p>
Identifier	803
Resource type	Corpus
URL	http://dcl.bas.bg/poscor/en/
Version	1.0
Last update	2011-11-20

Contacts

Angel Genov	
Position	Affiliated researcher
Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia dcltools@dcl.bas.bg

	http://dcl.bas.bg/PersonalPages/angel/
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia bulnc@dcl.bas.bg http://dcl.bas.bg
Availability start date	2011-11-20	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	http://dcl.bas.bg/poscor/

Metadata

Creation date	2011-11-20
Metadata last date updated	2013-01-31

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use
	NLP applications

Resource creation

--	--

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia bulnc@dcl.bas.bg http://ibl.bas.bg/en/
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds Own funds
	Start date	2011-02-01
	End date	2013-01-30
	Bulgarian National Corpus project	
	Project ID	BulNC
	URL	http://ibl.bas.bg/en/BGNC_en.htm
	Funding type	National funds
	Country	Bulgaria
	Start date	2009-12-17
	End date	2013-06-17

Resource documentation

Reports	Koeva, Svetla, Diana Blagoeva, Sia Kolkovska. Levels of annotation in the Bulgarian National Corpus. – Prace Filologiczne, 2012, LXIII, pp. 147-153. ISSN: 0138-0567. Koeva, Svetla, Svetlozara Leseva, Ivelina Stoyanova, Ekaterina Tarpomanova, Maria Todorova, Bulgarian Tagged Corpora. - In: Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages, 18-20 October 2006, Sofia, Bulgaria, 2006, pp. 78-86.
----------------	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
Modality	Modality type	Written language

Size	217,000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Sentence
	Lemmaization	
	Segmentation level	Word
	Morphosyntactic annotation – POS tagging	
	Segmentation level	Word

8.4. Bulgarian Sense-Annotated Corpus

General Information

Short name	BulSemCor
Description	The Bulgarian Sense-annotated Corpus (BulSemCor) contains sense-disambiguated lexical items defined in the context of occurrence. The Bulgarian Sense-annotated Corpus follows the methodology of the Princeton University SemCor. As BulSemCor it consists of excerpts from the Brown Corpus of Bulgarian. Each lexical item (simple word, compound word or multiword expression) is assigned manually the unique semantic or grammatical meaning from the Bulgarian wordnet (BulNet) in the particular context. Contrary to other sense annotated corpora, the BulSemCor covers both open and close class words and all occurrences of multiword expressions and named entities. The annotated lexical units inherit all the information from the synonym sets in the BulNet, incl. explanatory definition, PoS, usage examples, notes on grammatical, stylistic, and pragmatic properties, and all relations (semantic morpho-syntactic and extra-linguistic) pertaining to the synset, as well as the semantic and derivational relations pertaining to the literal. The BulSemCor contains 101 062 tokens, 99 480 annotated lexical units - 86 842 single words, a 5797 multiword expressions. The BulSemCor is used as training and testing set in the elaboration of a probability based automatic word-sense disambiguation that is applicable in variety of natural language processing tasks such as machine translation, text categorisation, information extraction, among others.
Identifier	804
Resource type	Corpus
URL	http://dcl.bas.bg/semcor/en/
Version	3.0
Last update	2011-11-20

Contacts

Angel Genov	
Position	Affiliated researcher

Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia dcltools@dcl.bas.bg http://dcl.bas.bg/PersonalPages/angel/
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia bulnc@dcl.bas.bg http://dcl.bas.bg
Availability start date	2010-11-30	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	http://dcl.bas.bg/semcor/

Metadata

Creation date	2011-11-20
Metadata last date updated	2013-01-31

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use

	NLP applications
--	------------------

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia bulnc@dcl.bas.bg http://ibl.bas.bg/
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30
	Bulgarian National Corpus project	
	Funding type	National funds
	Country	Bulgaria
	Start date	2009-12-17
	End date	2013-06-17

Resource documentation

Reports	<p>Koeva, Svetla, Svetlozara Leseva, Maria Todorova, Bulgarian Sense Tagged Corpus. - In: Proceedings of the 5th SALT MIL Conference on Minority Languages: Strategies for Developing Machine Translation for Minority Languages, Genoa, 2006, pp.79-87.</p> <p>Koeva, Svetla, Svetlozara Leseva, Ekaterina Tarpomanova, Borislav Rizov, Tsvetana Dimitrova and Hristina Kukova. Bulgarian Sense Annotated Corpus - Results and Achievements. In. Proceedings from the seventh international conference Formal Approaches to South Slavic and Balcan Languages, Dubrovnik, 2010, pp. 41-49. ISSN 978-953-55375-2-6</p>
----------------	--

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg

Modality	Modality type	Written language
Size	99000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Sentence
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation – POS tagging	
	Segmentation level	Word
	Semantic annotation – word senses	
	Segmentation level	Word

8.5. Bulgarian-X language Parallel Corpus

General Information

Short name	Bul-X-Cor
Description	<p>The Bulgarian-X language Parallel Corpus (Bul-X-Cor) is a part of the Bulgarian National Corpus (BulNC). The Bulgarian National Corpus is designed as a uniform framework for texts of different modality (written - spoken), period (synchronic - diachronic), and number of languages (monolingual - parallel where one of the counterparts is Bulgarian). Any X-language in the corpus is equally treated with respect to the text type diversity and balance, metadata description scheme, preprocessing and annotation, search engine queries and data storage format. Bulgarian-X Language Parallel Corpus includes parallel corpora of 48 languages – English, German, French, Slavic and Balkan languages, as well as other European and non-European languages. The parallel corpora represent only texts which have a Bulgarian correspondence – either the original is in Bulgarian, there is a Bulgarian translation, or both texts are translations from a third language. As of January 2013, the Bulgarian-X Language Parallel Corpus contains 4.2 billion tokens, comprising the biggest parallel corpus of Bulgarian. Languages are not equally represented: the largest parallel corpus is the Bulgarian-English parallel corpus (280.8 and 283.1 million words for Bulgarian and English respectively); there are 18 other corpora of over 200 million tokens per language, 2 parallel corpora between 100 and 200 million tokens per language, 11 parallel corpora of size in the range 5-15 million tokens per language, and the rest 15 are below 1 million, with the smallest corpus being Japanese with 50,000 tokens. Each parallel subcorpus within Bul-X-Cor mirrors the structure of BulNC. The structure, data formatting and text description follow the model of BulNC. All Bulgarian texts in BulNC and English texts in Bul-X-Cor are supplied with extensive metadata description compliant with the well established standards. The Bulgarian-English parallel corpus is supplied as well with annotation on various levels while the annotation of other languages has just started. Main applications of parallel corpora are in the field of computational linguistics: machine translation, developing bilingual lexical resources (dictionaries), etc. The benefits of the parallel corpora increase if they are annotated. The Bulgarian-X Language Parallel Corpus</p>

	Collocations service is a web service for collocations search and different types of statistics over the Bulgarian-X Language Parallel Corpus. The service employs the free of charge NoSketchEngine, a system for corpora processing that combines Manatee and Bonito. The Collocation service is a RESTful webservice, supporting complicated queries through http. Example: http://dcl.bas.bg/collocations/?cmd=collocations&word=нeтuser:bulncpass:bulnc The query returns the collocations of a given word in the NoSketchEngine format. The system also supports additional arguments, namely all that are accepted by NoSketchEngine, provided with default values and an optional language identifier. The following example restricts the statistics to Bulgarian: http://dcl.bas.bg/collocations/?cmd=collocations&word=нeт&lang=bg
Identifier	805
Resource type	Corpus
URL	http://www.ibl.bas.bg/en/BGNC_parallel_en.htm http://search.dcl.bas.bg/en/
Version	2.0
Last update	2012-07-20

Contacts

Angel Genov	
Position	Affiliated researcher
Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia dcltools@dcl.bas.bg http://dcl.bas.bg/PersonalPages/angel/
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia bulnc@dcl.bas.bg http://dcl.bas.bg
Availability start date	2011-09-01	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	http://search.dcl.bas.bg
Restrictions of use	Academic – non-commercial use
Access medium	Web executable
Execution location	http://dcl.bas.bg/collocations/?cmd=collocations&word=HEТ

Metadata

Creation date	2011-11-20
Metadata last date updated	2013-02-01

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30
	Bulgarian National Corpus project	

Project short name	BulNC
URL	http://ibl.bas.bg/en/BGNC_en.htm
Funding type	National funds
Start date	2009-12-17
End date	2013-06-17

Resource documentation

Reports	<p>Koeva, Svetla, Ivelina Stoyanova, Rositsa Dekova, Borislav Rizov, Angel Genov. Bulgarian X-language Parallel Corpus. – In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul: European Language Resources Association (ELRA), 2012, pp. 51-62. ISBN: 978-2-9517408-7-7.</p> <p>Koeva, Svetla, Ivelina Stoyanova, Svetlozara Leseva, Tsvetana Dimitrova, Rositsa Dekova, Ekaterina Tarpomanova. The Bulgarian National Corpus: Theory and Practice in Corpus Design. – Journal of Language Modelling, 2012, 1 (1), pp. 65-110. ISSN: 2299-8470 http://nlp.ipipan.waw.pl/ojs/index.php/JLM/issue/current</p>
Tool documentation type	<p>Help functions</p> <p>Manual</p> <p>None</p>

Texts

Media type	text	
Linguality type	Multilingual	
Multilinguality type	Parallel	
Languages	English	
	Language ID	en
	Size	260,681,821 tokens
	Romanian	
	Language ID	ro
	Size	235,859,637 tokens
	Greek	
	Language ID	el
	Size	229,749,068 tokens
	Czech	
	Language ID	cs
	Size	196,769,297 tokens
	Polish	
	Language ID	pl

Size	197,762,449 tokens
Slovak	
Language ID	sk
Size	189,752,630 tokens
Spanish	
Language ID	es
Size	191,092,782 tokens
Danish	
Language ID	da
Size	190,843,358 tokens
Finnish	
Language ID	fi
Size	156,288,741 tokens
Hungarian	
Language ID	hu
Size	183,530,929 tokens
Estonian	
Language ID	et
Size	160,175,247 tokens
Slovene	
Language ID	sl
Size	188,776,967 tokens
German	
Language ID	de
Size	194,497,872 tokens
Lithuanian	
Language ID	lt
Size	170,381,570 tokens
Italian	
Language ID	it
Size	209,083,677 tokens
Bosnian	
Language ID	bs
Size	6,195,646 tokens
Galician	

Language ID	ga
Size	629,272 tokens
Croatian	
Language ID	hr
Size	11,950,183 tokens
Latvian	
Language ID	lv
Size	167,600,804 tokens
Macedonian	
Language ID	mk
Size	9,542,940 tokens
Maltese	
Language ID	mt
Size	163,515,445 tokens
Dutch	
Language ID	nl
Size	204,309,755 tokens
Portuguese	
Language ID	pt
Size	211,824,204 tokens
Albanian	
Language ID	sq
Size	9,781,443 tokens
Swedish	
Language ID	sv
Size	180,752,058 tokens
Turkish	
Language ID	tr
Size	13,297,328 tokens
Arabic	
Language ID	ar
Size	2,446,857 tokens
Azerbaijani	
Language ID	az
Size	137,238 tokens

Catalan; Valencian	
Language ID	ca
Size	640,522 tokens
Basque	
Language ID	eu
Size	461,080 tokens
French	
Language ID	fr
Size	231,486,663 tokens
Irish	
Language ID	ga
Size	13,287,693 tokens
Hebrew	
Language ID	he
Size	2,872,765 tokens
Armenian	
Language ID	hy
Size	139,802 tokens
Icelandic	
Language ID	is
Size	762,894 tokens
Japanese	
Language ID	ja
Size	50,194 tokens
Georgian	
Language ID	ka
Size	128,502 tokens
Kazakh	
Language ID	kk
Size	486,766 tokens
Kirghiz; Kyrgyz	
Language ID	ky
Size	135,031 tokens
Mongolian	
Language ID	mn

	Size	135,076 tokens
	Norwegian	
	Language ID	no
	Size	1,588,561 tokens
	Russian	
	Language ID	ru
	Size	3,293,243 tokens
	Serbian	
	Language ID	sr
	Size	1,832,323 tokens
	Tajik	
	Language ID	tg
	Size	160,123 tokens
	Turkmen	
	Language ID	tk
	Size	127,430 tokens
	Ukrainian	
	Language ID	uk
	Size	744,815 tokens
	Chinese	
	Language ID	zh
	Size	229,293 tokens
	Size	4,195,791,994 tokens
	Character encoding	UTF-8
	Annotation	Alignment
		Segmentation level Sentence
		Segmentation
		Segmentation level Sentence
		Segmentation
		Segmentation level Word
		Lemmatization
		Segmentation level Word
		Morphosyntactic annotation - below POS tagging

	Segmentation level	Word
Media type	text	
Linguality type	Multilingual	
Multilinguality type	Parallel	
Languages	Bulgarian	
	Language ID	bg
Size	1,202,209,147 Tokens tokens	

8.6. Bulgarian WordNet

General Information

Short name	BulNet
Description	<p>The Bulgarian wordnet (BulNet) is a lexical-semantic network of Bulgarian that represents lexical knowledge in the form of interconnected nodes. It is an electronic thesaurus with a structure modelled on the structure of the Princeton WordNet (PWN) that, subsequently, was followed by the EuroWordNet and BalkaNet. The Bulgarian wordnet describes the meaning of a lexical unit by placing it within a network of semantic relations, such as hypernymy, meronymy, antonymy, among others. It is one of the most complete and consistent lexical resources as its literals are much more than the units in any the word list in a standard spelling dictionary. The synonym sets pertaining to different languages are connected by means of inter-language equivalence relations, which are used for the development of the wordnet multilingual lexical-semantic network in the global wordnet. The Bulgarian wordnet is one of the biggest in Europe. It contains 49,189 synonym sets (as of January 21, 2013) distributed into nine parts of speech - nouns, verbs, adjectives, adverbs (open-class words); pronouns, prepositions, conjunctions, particles, interjections, closed-class words). Each synonym set is supplied with explanatory definition which represents the common referential meaning of all its members. The Bulgarian WordNet is a language-internal structure, minimally containing: set of variants or synonyms making up the synset; part-of-speech; language-internal relations to other synsets; a unique-id linking the synset to the English Wordnet 3.0. Number of Synonym sets: 49,189 Number of Literals: 103,506 Number of Relations: 116,025 Number of Semantic relations: 70,317 Number of Extralinguistic relations: 5648 The Bulgarian WordNet is distributed without: glosses, usage labels, morpho-syntactic properties, examples.</p>
Identifier	806
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Wordnet
URL	http://dcl.bas.bg/en/wordnet_en.html
Version	5.0
Last update	2013-01-20

Contacts

--

Svetla Koeva	
Position	Professor
Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia svetla@dcl.bas.bg http://dcl.bas.bg/PersonalPages/svetla/svetla.html

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia bulnet@dcl.bas.bg http://dcl.bas.bg
Availability start date	2004-12-01	

Licences

ELRA_END_USER		
Restrictions of use	Academic – non-commercial use	
Access medium	CD-ROM	
Fee	7,114.50 euro	
User nature	Commercial	
Membership	Member	False
	Membership institution	ELRA
ELRA_VAR		
Restrictions of use	Commercial use	
Fee	5,928.75 euro	
User nature	Academic	
Membership	Member	True
	Membership institution	ELRA
ELRA_END_USER		
Restrictions of use	Academic – non-commercial use	

Access medium	CD-ROM	
Fee	237.15 euro	
User nature	Academic	
Membership	Member	True
	Membership institution	ELRA
ELRA_END_USER		
Restrictions of use	Academic – non-commercial use	
Access medium	CD-ROM	
Fee	3,557.25 euro	
User nature	Commercial	
Membership	Member	True
	Membership institution	ELRA
ELRA_VAR		
Restrictions of use	Commercial use	
Access medium	CD-ROM	
Fee	5,928.75 euro	
User nature	Academic	
Membership	Member	True
	Membership institution	ELRA
ELRA_VAR		
Restrictions of use	Commercial use	
Fee	11,857.50 euro	
User nature	Commercial	
Membership	Member	False
	Membership institution	ELRA
ELRA_VAR		
Restrictions of use	Commercial use	
Fee	11,857.50 euro	
User nature	Academic	
Membership	Member	False
	Membership institution	ELRA
ELRA_END_USER		

Restrictions of use	Academic – non-commercial use	
Fee	474.30 euro	
User nature	Academic	
Membership	Member	False
	Membership institution	ELRA

Metadata

Creation date	2013-01-29
Metadata last date updated	2013-01-31

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use
	NLP applications

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/en/index.htm
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30

BalkaNet - Design and Development of a Multilingual Balkan WordNet	
Project ID	BalkaNet
URL	http://www.dblab.upatras.gr/balkanet/
Funding type	EU funds
Start date	2001-09-01
End date	2004-08-31
BulNet - A Lexical-Semantic Network of Bulgarian	
Project short name	BulNet
Project ID	BulNet
URL	http://dcl.bas.bg/BulNet/general_en.html
Funding type	National funds
Creation start date	2001-09-01

Resource documentation

Reports	<p>Koeva, Svetla. Bulgarian Wordnet - current state, applications and prospects. - In: Bulgarian-American Dialogues. Sofia: Prof. M. Drinov Academic Publishing House. 2010, pp. 120-132. ISBN 978-954-322-383-1</p> <p>Koeva, Svetla. Derivational and Morphosemantic Relations in Bulgarian Wordnet. - In: Intelligent Information Systems, XVI, Warsaw, Academic Publishing House, 2008, pp. 359-389. ISBN 978-93-60434-44-4</p> <p>Stamou, S., K. Ofizer, K. Pala, D. Christodoulakis, D. Cristea, D. Tufis, S. Koeva, G. Totkov, D. Dutoit, and M. Grigoriadou. A multilingual semantic network for Balkan languages, In Proceedings of the First Global WordNet Conference (GWC), Mysore, India, 2002.</p> <p>Koeva, S., G. Totkov and A. Genov. Towards Bulgarian WordNet. Romanian Journal of Information Science and Technology, Vol. 7, No. 1-2, 45-61, 2004.</p> <p>Koeva, S., T. Tinchev and S. Mihov. Bulgarian Wordnet - structure and validation. In Romanian Journal of Information Science and Technology, Vol. 7, No. 1-2, 61-78, 2004.</p> <p>Krstev, C., S. Koeva and D. Vitas. Towards the Global Wordnet. In First International Conference of Digital Humanities Organizations (ADHO) Digital Humanities 2006, Paris-Sorbonne, 2006, 114-117.</p> <p>Koeva, S. Derivational and morphosemantic relations in Bulgarian Wordnet. - In: Intelligent Information Systems, XVI, Warsaw, Academic Publishing House, 2008, 359-389.</p> <p>Koeva, S., C. Krstev, and D. Vitas. Morpho-semantic relations in Wordnet - a case study for two Slavic languages. In Proceedings of the Fourth Global WordNet Conference, Szeged, 2008, 239-254.</p> <p>Koeva, S. Bulgarian Wordnet - current state, applications and prospects. - In: Bulgarian-American Dialogues, Prof. M. Drinov Academic Publishing House, Sofia, 2010, 120-132.</p>
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Wordnet
---	---------

Lexical conceptual resource encoding	Encoding level	Semantics
	Linguistic information	Part of speech Semantics – cross references Semantics – domain Semantics – relations Semantics – relations – antonyms Semantics – relations – hyperonyms Semantics – relations – hyponyms Semantics – relations – meronyms Semantics – relations – synonyms
	Conformance to standards best practices	Word net

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
	Language script	Cyrillic
Modality	Modality type	Written language
Size	49,189 synsets	
Character encoding	UTF-8	

8.7. Bulgarian WordNet - web access

General Information

Short name	WordNetWeb
Description	<p>The Bulgarian WordNet is a network of lexical-semantic relations, an electronic thesaurus with a structure modelled on that of the Princeton WordNet and those constructed in the EuroWordNet and BalkaNet projects. Wordnet service is an online service that gives the users access to a subset of the Bulgarian wordnet (BulNet), containing over 49,000 synonym sets (synsets), as of January 2013, and to the entire database of the Princeton Wordnet (PWN). The system is RESTful webservice and supports two sorts of queries through http. 1. Search for objects where the query described in the WordNet modal language returns a list of object identifiers for which it is true. 2. Search for information about objects and returns a list of data for: Literal: identifier, word, lemma. Synset: identifier, ili, POS, definition, stamp, bcs, language (identifier), frequency. Note: identifier, text. Example: http://dcl.bas.bg/wn/?cmd=query&query=word('дума') user: bulnetpass: bulnet</p> <p>The two sorts of queries support nonobligatory parameter (format) showing the type of the result. If the value of the format is json, the result is coded as json, otherwise it is not coded. Users can search for synonyms, hypernyms, antonyms, and translation equivalents of</p>

	different words and lemmas in the following language pairs: English-English, English-Bulgarian, Bulgarian-English, and Bulgarian-Bulgarian.
Identifier	807
Resource type	Tool/service
Tool/service type	Service
URL	http://dcl.bas.bg/BulNet/general_en.html
Version	1.0
Last update	2011-11-20

Contacts

Angel Genov	
Position	Affiliated researcher
Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia dcltools@dcl.bas.bg http://dcl.bas.bg/PersonalPages/angel/
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia bulnet@dcl.bas.bg http://dcl.bas.bg
Availability start date	2011-11-20	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Web executable
Execution location	http://dcl.bas.bg/wn/?cmd=query&query=word('дума')

Metadata

Creation date	2011-11-20
Metadata last date updated	2013-02-01

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use

Resource creation

Resource creator	Borislav Rizov	
	Position	Assistant Professor
	Contact	boby@dcl.bas.bg
	Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds Own funds
	Start date	2011-02-01
	End date	2013-01-30
	BalkaNet - Design and Development of a Multilingual Balkan WordNet	
	Project short name	BalkaNet
	URL	http://www.dblab.upatras.gr/balkanet/
	Funding type	EU funds
	Start date	2001-09-01
	End date	2004-08-31
	BulNet - A Lexical-Semantic Network of Bulgarian	
	Funding type	National funds

Resource documentation

Reports	<p>Koeva Sv. Bulgarian Wordnet - current state, applications and prospects, In: Bulgarian-American Dialogues, Prof. M. Drinov Academic Publishing House Sofia, 120-132, 2010. ISBN 978-954-322-383-1</p> <p>Koeva, Sv. Derivational and Morphosemantic Relations in Bulgarian Wordnet. In: Intelligent Information Systems, XVI, Warsaw, Academic Publishing House, pp. 359—389, 2008. ISBN 978-93-60434-44-4</p> <p>Rizov, Borislav. Processing Wordnet with Modal Logic, Tadić et al. (eds) Proceedings of the 6th International Conference on Formal Approaches to South Slavic and Balkan Languages, 25—28 September 2008, Dubrovnik, pp. 93-100.</p>
Tool documentation type	<p>Help functions</p> <p>Online</p>

Tool/service

Tool/service type	Service			
Language dependent	False			
Input	Media type	text		
Output	Media type	text		
Operating system	OS-independent			
Tool/service evaluation	Evaluated	True		
	Level	Diagnostic		
	Evaluators	Borislav Rizov		
		Position	Assistant Professor	
		Contact	boby@dcl.bas.bg http://dcl.bas.bg/en/people_en/boby.html	

8.8. Bulgarian Spell Checker for Windows

General Information

Short name	WinEst
Description	<p>The system for automatic spelling checking WinEst for Microsoft Office detects and marks the incorrectly written words in a text and suggests the most probable candidates to correct the errors. WinEst offers the entire potential of the contemporary spelling correction: proficiently compiled dictionary, which contains over a million and a half words, and replacement suggestions, which are ordered according to their probability. WinEst is based on the Electronic Grammar Dictionary of Bulgarian, developed at the Department of Computational Linguistics, which contains over 85 000 words. It contains logic for detection of careless mistakes (wrong key pressed, letter swapping, skipped letters or extra letters), identifies errors of ignorance and integrates perfectly into the dictionaries used in Microsoft Office. WinEst uses an extremely fast and effective method for searching and detecting the correct words regardless of the text size. The functionality of the product is realized through the use of minimal acyclic deterministic automata and Levenshtein automata, which allow maximum speed, precision and coverage. A distinctive feature of</p>

	WinEst is it is easy to install and uninstall, and no System restart is required. Advantages: WinEst offers the entire potential of the contemporary spelling checking and correction. Together with the proficiently compiled dictionary the product is capable of finding replacement suggestions, which are ranked by probability. Representativeness: covers the basic wordstock of Bulgarian. Precision: all words are checked by experts. Convenience: the replacement candidates are ranked by probability. A module for Cyrillic layout: WinEst works perfectly both with the standard BDS layout and with the various phonetic layouts. WinEst is a 32-bit module and thus requires a 32-bit Microsoft Office.
Identifier	808
Resource type	Tool/service
Tool/service type	Tool
URL	http://dcl.bas.bg/est/index_en.php http://dcl.bas.bg/en/winest.html
Version	2.0
Last update	2011-11-20

Contacts

Angel Genov	
Position	Affiliated researcher
Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia dcltools@dcl.bas.bg http://dcl.bas.bg/PersonalPages/angel/
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use
Availability start date	2011-06-01

Licences

CC-BY-ND	
Restrictions of use	No redistribution
Access medium	Downloadable
Download location	http://dcl.bas.bg/sites/default/files/webfm/WinEst/winestSetup.exe

Metadata

--	--

Creation date	2011-11-20
Metadata last date updated	2013-02-01

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use
Actual uses	Human use

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds Own funds
	Start date	2011-02-01
	End date	2013-01-30
	Web applications for editing Bulgarian texts	
	Funding type	National funds
Creation start date	2010-01-01	

Resource documentation

Reports	Oliva, Karel, Svetla Koeva. Sintaksis na nevazmozhnoto (Syntax of the Impossible). - Balgarski ezik, 2009, 3, pp. 7-17. ISSN 0005-4283. (in Bulgarian)
----------------	--

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Bulgarian	
	Media type	text
	Language ID	bg
Output	Media type	text
Operating system	Windows	

Tool/service evaluation	Evaluated	True
	Level	Diagnostic

8.9. Bulgarian Spell Checker Web Service

General Information

Short name	WebEst
Description	The development of web based applications assisting the work with Bulgarian texts is imposed, on the one hand, by the wider use of the Internet in everyday communications of various types (work, education, administration, media), and on the other hand, by the lack of modern web based linguistic applications (for Bulgarian). The creation of modern web based linguistic applications (web services, web components and web applications) which offer a possibility for effective work with no respect to operation systems, text processing applications or browsers. The Spell Checker is integrated as a web service – both the web service integration and the online spelling checking (as an illustration of the integration) are possible. The Spell Checker is based on the construction of a dictionary in a minimal acyclic deterministic automaton and offers replacement suggestions on the basis of Levenshtein automata. WebEst allows the users to check and correct Bulgarian texts on the Internet. The Spell Checker web service can be used in different blogs, chat forums, online shops, media, and everywhere in the creation of Internet contents, so that it will assist the correct writing of Bulgarian texts. The advantages of the web based linguistic applications can be summarized as follows: they are more accessible to use as they are not related to any operation system or web browser. The wider use of the Internet not only as an environment for communication but also as an operating environment, which includes text creation and editing, increases the importance of the project outcomes.
Identifier	809
Resource type	Tool/service
Tool/service type	Service
URL	http://dcl.bas.bg/est/index_en.php
Version	2.0
Last update	2011-11-20

Contacts

Angel Genov	
Position	Affiliated researcher
Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia dcltools@dcl.bas.bg http://dcl.bas.bg/PersonalPages/angel/
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use
Availability start date	2011-06-01

Licences

CC-BY	
Restrictions of use	Other
Access medium	Web executable
Execution location	http://dcl.bas.bg/est/index_en.php#tabs-5 http://dcl.bas.bg/est/checker.php

Metadata

Creation date	2011-11-20
Metadata last date updated	2013-02-01

Usage

Foreseen use	Human use
Actual uses	Human use

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd. 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/en/index.htm
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds Own funds
	Start date	2011-02-01

	End date	2013-01-30
	Web applications for editing Bulgarian texts	
	Funding type	National funds
Creation start date	2010-01-01	

Resource documentation

Reports	Oliva, Karel, Svetla Koeva. Sintaksis na nevazmozhnoto (Syntax of the Impossible). - Balgarski ezik, 2009, 3, pp. 7-17. ISSN 0005-4283. (in Bulgarian)
----------------	--

Tool/service

Tool/service type	Service	
Language dependent	False	
Input	Bulgarian	
	Media type	text
	Language ID	bg
Output	Media type	text
Operating system	OS-independent	
Tool/service evaluation	Evaluated	True
	Level	Diagnostic

8.10. The Bulgarian-X Language Parallel Corpus Collocations service

General Information

Description	<p>The Bulgarian-X Language Parallel Corpus Collocations service is a web service for collocations search and different types of statistics over the Bulgarian-X Language Parallel Corpus.</p> <p>The service employs the free of charge NoSketchEngine, a system for corpora processing that combines Manatee and Bonito.</p> <p>The Collocation service is a RESTful webservice, supporting complicated queries through http.</p> <p>user: bulnc pass: bulnc</p> <p>The query returns the collocations of a given word in the NoSketchEngine format.</p> <p>The system also supports additional arguments, namely all that are accepted by NoSketchEngine, provided with default values and an optional language identifier.</p> <p>The Collocation service is one of the 4 different means of access to the Bulgarian-X Language Parallel Corpus.</p>
--------------------	--

8.11. Lists of Bulgarian Multiword Expressions

General Information

Short name	BulMWEs
Description	The classification of multiword expressions (MWEs) developed by Baldwin et al. (Baldwin, T., C. Bannard, T. Tanaka, D. Widdows. An Empirical Model of Multiword Expression Decomposability. In: Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment. 2003) who distinguish between non-decomposable, idiosyncratically decomposable and simple decomposable MWEs is adopted. Further, we divide simple decomposable MWEs into 10 categories based on pragmatic factors – whether they are or contain a named entity (NE). Free collocations are free phrases (non-MWEs) which are statistically marked, i.e. appear with high frequency in a corpus, but are not linguistically marked. The lists of Multiword expressions are the result of automatic and semi-automatic tagging and classification of the corpus Wiki1000+ (13.4 million tokens): Non-decomposable - 700, Idiosyncratically decomposable - 3,156, Simple decomposable (NEs without connection between elements - 36,932, NEs with a meaningful element(s) - 11,248, Non-NEs with a vague connection between components - 1,46, NEs with meaningful components but connection difficult to restore - 1,086, NEs with descriptor and additional element - 18,962, Non-NEs with a NE as one of the components - 27,373, Non-NEs with a standard, easy to restore connection between components- 140,394, NEs with a standard, easy to restore connection between components - 16,653, Non-NEs with explicit connection between components - 1,468), “Free collocations” - 49,651, Free phrases- 1,197,762.
Identifier	811
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://dcl.bas.bg/en/dictionaries_en.html
Version	1.0
Last update	2012-07-20

Contacts

Ivelina Stoyanova	
Position	Affiliated researcher
Contact	iva@dcl.bas.bg http://dcl.bas.bg/en/people_en/iva.html
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use
IPR holder	Institute for Bulgarian Language

	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia dcl@dcl.bas.bg http://dcl.bas.bg
Availability start date	2012-04-01	

Licences

Restrictions of use	Academic – non-commercial use
Download location	http://dcl.bas.bg/Resources/MWEs/lists.zip
Fee	free of charge

Metadata

Creation date	2012-07-27
Metadata last date updated	2013-01-31

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Named entity recognition	
Actual uses	Human use	
	NLP applications	
	NLP-specific use	Named entity recognition

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17

		1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/en/index.htm
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds Own funds
	Start date	2011-02-01
	End date	2013-01-30
Creation start date	2012-01-01	

Resource documentation

Reports	Stoyanova, Ivelina. PhD thesis: Automatic recognition and annotation of compound lexical units in Bulgarian (in Bulgarian). Lists of MWE of different categories (Classification 6, p. 76)
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
	Language script	Cyrilic
Modality	Modality type	Written language
Size	27784 multi-word units	
Character encoding	UTF-8	

8.12. Bulgarian Frequency Dictionaries

General Information

--	--	--

Short name	BulFreq
Description	Bulgarian Frequency Dictionaries are lemma frequency dictionaries extracted from the Bulgarian National Corpus (BulNC) which is annotated at various linguistic levels - sentence segmentation, POS tagging, lemmatisation, etc. BulNC contains 6 domain-specific subcorpora and thus 6 domain-specific Freq dictionaries were developed independently, as well as a general dictionary which combines all domain-specific ones. Each dictionary is available in 2 versions: in alphabetical order and in frequency order. Frequencies are automatically collected; more efficient methods for compilation of frequency lists and dictionaries are still investigated. The compilation of a frequency dictionary is performed in stages – compilation of the dictionary on smaller parts of the corpus, followed by merging.
Identifier	812
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://dcl.bas.bg/en/frequency_en.html
Version	1.0
Last update	2012-07-20

Contacts

Angel Genov	
Position	Affiliated researcher
Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia dcltools@dcl.bas.bg http://dcl.bas.bg/PersonalPages/angel/
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia dcl@dcl.bas.bg http://dcl.bas.bg
Availability start date	2012-04-01	

Licences

CC-BY-NC	
Restrictions of use	Academic – non-commercial use
Access medium	Downloadable
Download location	http://dcl.bas.bg/Resources/Frequency/Frequency.zip

Metadata

Creation date	2012-07-27
Metadata last date updated	2013-01-31

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use
	NLP applications

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/en/index.htm
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds Own funds
	Start date	2011-02-01
	End date	2013-01-30

Creation start date	2012-01-01
----------------------------	------------

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary
---	-----------------------------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
	Language script	Cyrilic
Modality	Modality type	Written language
Size	2142555 words	
Character encoding	UTF-8	

8.13. Hydra - tool for developing wordnets

General Information

Short name	Hydra
Description	Hydra is a tool for editing, viewing, searching and validating wordnet. The Hydra API for wordnet processing uses abstract language independent of the data representation, the tool supports a multiple-user concurrent access for editing and browsing arbitrary number of monolingual wordnets, it optimizes data visualization as well as enhances editing, undo/redo functions, etc. The search engine works with the wordnet modal language. The language abstracts the internal data representation and is expressive for the most of the tasks in processing wordnets. Provided that a given wordnet property is definable as a formula in the modal language, the tool determines all the objects in the wordnet structure validating the formula, and hence the property, covering an automatic consistency validation. As a platform-independent system, Hydra has been successfully tested under Linux and Windows.
Identifier	813
Resource type	Tool/service
Tool/service type	Tool
URL	http://dcl.bas.bg/en/hydra.html
Version	0.6
Last update	2012-07-20

Contacts

--

Borislav Rizov	
Position	Assistant Professor
Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia boby@dcl.bas.bg
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg
Availability start date	2005-06-01	

Licences

GPL	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://dcl.bas.bg/Tools/Hydra/hydra.zip

Metadata

Creation date	2013-01-29
Metadata last date updated	2013-02-01

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use Human use
---------------------	------------------------

Resource creation

Resource creator	Borislav Rizov	
	Position	Assistant Professor
	Contact	boby@dcl.bas.bg
	Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30
Creation start date	2010-01-01	

Resource documentation

Reports	Rizov, Borislav. Hydra: A Modal Logic Tool for Wordnet Development, Validation and Exploration. - In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, 2008, European Language Resources Association (ELRA) electronic publication. ISBN 2-9517408-4-0	
Documents	Manual	
	Title	Hydra Installation Manual
	URL	http://dcl.bas.bg/Tools/Hydra/Hydra-InstallationManual.pdf
	Manual	
	Title	Hydra User Manual
	URL	http://dcl.bas.bg/Tools/Hydra/Hydra-UserManual.pdf
Tool documentation type	Manual Online	

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Media type	text
Output	Media type	text
Operating system	OS-independent	

Tool/service evaluation	Evaluated	True	
	Level	Diagnostic	
	Evaluators	Borislav Rizov	
		Position	Assistant Professor
		Contact	boby@dcl.bas.bg http://dcl.bas.bg/en/people_en/boby.html
		Organization	Institute for Bulgarian Language Department of Computational Linguistics est@dcl.bas.bg
Tool/service creation	Implementation language	Python	

8.14. Chooser - annotation tool

General Information

Short name	Chooser
Description	Chooser is an OS independent multi-functional system for linguistic annotation, adaptable to different annotation schemata. The basic annotation functionalities of the tool are: (i) fast and easy-to-perform selection; (ii) run-time access to information for the candidate senses such as definition, frequency, the associated wordnet synsets with all the pertaining info – synonyms, gloss, semantic relations, notes on usage, form, etc.; (iii) identification of MWEs with contiguous and non-contiguous constituents and supplying information for them at run-time. The basic functions are enhanced with flexible text navigation strategies - forward and backward navigation over: (i) all words; (ii) non-annotated words; (iii) all instances of a word; (iv) all instances of a sense. Finally, a flexible search strategy allowing both exact match search according to word form or lemma, and regular expression search is integrated. The tool interface features a fully-fledged visualization of the wordnet synsets for the candidate senses available for a selected LU through coupling with the system for wordnet development and exploration Hydra. A unified wordnet representation in Chooser and Hydra is implemented. Chooser provides multiple-user concurrent access and dynamic real-time update in the knowledge base, so that all changes, such as newly-encoded synsets, literals, relations, are updated in both systems and made available to all the users immediately.
Identifier	814
Resource type	Tool/service
Tool/service type	Tool
URL	http://dcl.bas.bg/en/Chooser.html
Version	3.0
Last update	2012-07-20

Contacts

Borislav Rizov

Position	Assistant Professor
Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia boby@dcl.bas.bg
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia est@dcl.bas.bg http://dcl.bas.bg
Availability start date	2003-06-01	

Licences

GPL	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://dcl.bas.bg/Tools/Chooser/chooser.zip

Metadata

Creation date	2013-01-29
Metadata last date updated	2013-02-01

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use
Actual uses	Human use

Resource creation

Resource creator	Borislav Rizov	
	Position	Assistant Professor
	Contact	boby@dcl.bas.bg
	Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30
Creation start date	2010-01-01	

Resource documentation

Reports	Koeva, Svetla, Borislav Rizov, Svetlozara Leseva. Chooser - A Multi-task Annotation Tool. - In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, European Language Resources Association (ELRA) electronic publication, 2008, pp. 728-734. ISBN 2-9517408-4-0	
Documents	Manual	
	Title	Chooser Installation Manual
	URL	http://dcl.bas.bg/Tools/Chooser/Chooser-InstallationManual.pdf
	Manual	
	Title	Chooser User Manual
	URL	http://dcl.bas.bg/Tools/Chooser/Chooser-UserManual.pdf
Tool documentation type	Manual Online	

Tool/service

Tool/service type	Tool	
Language dependent	False	
Input	Media type	text
Output	Media type	text
Operating system	OS-independent	
Tool/service	Evaluated	True

evaluation	Level	Diagnostic	
	Evaluators	Borislav Rizov	
		Position	Assistant Professor
		Contact	boby@dcl.bas.bg http://dcl.bas.bg/en/people_en/boby.html
		Organization	Institute for Bulgarian Language Department of Computational Linguistics est@dcl.bas.bg
Tool/service creation	Implementation language	Python	

8.15. Bulgarian Sentence Splitter and Tokenizer

General Information

Short name	BulSST
Description	The sentence splitter marks the sentence boundaries and the tokenizer marks string of symbols in raw Bulgarian text. The sentence splitter applies regular rules and lexicons. Both - regular rules and lexicons - are manually crafted by an expert. Lists of lexicons (for recognizing abbreviations after which there must be or there might be a capital letter, a number, etc. in the middle of the sentence) are applied before the regular rules. The lexicons are compiled by a separate tool - the Lexicon compiler, as minimal acyclic final state automata which allows an effective processing. Sentence borders are represented as a position and length which allows the incoming text to be kept unchanged as well as an easy integration in different systems for annotation. The tokenizer demarcates strings of letters, numbers, punctuation marks, special symbols, combinations of them and empty symbols. Regular patterns are used to recognize some simple cases of named entities that mean dates, fractions, emails, internet addresses, abbreviations, etc. The tokenizer classifies each recognized token (for example: small Cyrillic letters, capital Latin letters, etc.). The tokenizer utilizes finite state transducers for token recognition and type matching.
Identifier	815
Resource type	Tool/service
Tool/service type	Tool
URL	http://dcl.bas.bg/en/programs_en.html #BGTokenizer
Version	3.0
Last update	2012-07-20

Contacts

Angel Genov	
Position	Affiliated researcher
Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia

	dcltools@dcl.bas.bg http://dcl.bas.bg/PersonalPages/angel/
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia est@dcl.bas.bg http://dcl.bas.bg
Availability start date	2005-06-01	

Licences

CC-BY-NC	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://dcl.bas.bg/Tools/TokenizerSplitter/TokenizerSplitter-linux32.zip http://dcl.bas.bg/Tools/TokenizerSplitter/TokenizerSplitter-linux64.zip

Metadata

Creation date	2012-07-20
Metadata last date updated	2013-01-31

Validation

Validated	True
------------------	------

Usage

Foreseen use	NLP applications
Actual uses	NLP applications

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds Own funds
	Start date	2011-02-01
	End date	2013-01-30
Creation start date	2010-01-01	

Resource documentation

Reports	Koeva, Svetla, Angel Genov. Bulgarian language processing chain. In Proceedings of Integration of Multilingual Resources and Tools in Web Applications. Proceedings of a Workshop in conjunction with GSCL 2011, University of Hamburg, 2011.
----------------	---

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Bulgarian	
	Media type	text
	Language ID	bg
	Segmentation level	Sentence Word
Output	Media type	text
Operating system	Linux	
Tool/service evaluation	Evaluated	True
	Level	Diagnostic

8.16. Web based infrastructure for Bulgarian data processing

General Information

Short name	DCLservices
Description	<p>The Bulgarian Language Processing Chain includes the following types of text processing and linguistic annotation: Sentence segmentation; Tokenisation; POS tagging and grammatical annotation; Lemmatisation. The Bulgarian POS tagger marks up each word with the most probable Part of Speech and unambiguous morphosyntactic information among the set of tags associated with a given word. The tagger is based on SVM (Support Vector Machines) learning. The tagger predicts the POS tag of a word based on a set of features describing the word and its context. These features are words, word bigrams and trigrams within a window of words around the currently tagged word; POS tags, POS tags bigrams and trigrams in the current window, and information about suffixes, prefixes, capitalization, hyphenation etc. for the unknown words. The tagger is trained and tested on manually POS disambiguated corpus. The strategy chosen for training Bulgarian tagger is two passes in both directions; a window of five tokens, the currently tagged word being on the second position; two and three-grams of words or tags or ambiguity classes, lexical parameters as prefixes, suffixes, sentence borders, and capital letters. The trained model is applied to disambiguate texts. The precision of the tagger up to the moment is 96,58%. The Bulgarian lemmatizer determines for a given word form its lemma and detailed morphosyntactic annotation. The lemmatization is based on an unambiguous association between the tagger output and information encoded in a large grammatical dictionary of Bulgarian language. At the tagging a reduced tagset is used (75 word classes compering to 1029 unique grammatical tags in the dictionary) compiled in a way that the minimum necessary information for unambiguous association with the respective lemma to be ensured. A small number of rules and preferences are also implemented to limit the ambiguity in lemmatization. Some additional tools for advanced processing and annotation are available, as well as for annotation and alignment of parallel texts at sentential and subsentential level. A highly scalable web service based infrastructure was developed to provide easy access to the tools for text processing and annotation of Bulgarian. Three different types of access is provided to facilitate the user access to the system: online access; access via RESTful API; asynchronous access. Online access is suitable for users who need processing of relatively small amount of data occasionally. RESTful API access is suitable for software developers who can integrate the processing tools in high level applications. Asynchronous access is aimed for processing large corpora – the user uploads the archived corpus, it is processed on the server, a notification email is sent upon completion of the task and the annotated corpus can be downloaded. The system is highly scalable and can be distributed on different machines. The service infrastructure consist of three main components: Frontend, Backend and TaskDispatcher, each of these can be deployed on different machines. The Frontend component is responsible for implementation of the access policies of the service apis, error handling, logging, support of different return formats (xml,json,plain text), communication with the Backend. Also the Fronted provides the Web UI to user to control the asynchronous tasks: start, stop or monitor a task and upload/download data. The Backend performs the actual processing and it combines the Bulgarian tokenizer, sentence splitter, tagger and lemmatiser in the form of a server application which handles the requests of the Frontend over tcp/ip. Even though the Frontend is implemented efficiently and can handle many request simultaneously, whenever necessary several instances of the Frontend can be distributed on different machines. The TaksDispatcher is responsible for managing the processes of the asynchronous tasks. It receives the start/stop commands by the Frontend and notifies the user by e-mail when the result is ready.</p>
Identifier	816
Resource type	Tool/service

Tool/service type	Tool
URL	http://dcl.bas.bg/en/DCLservices.html http://dcl.bas.bg/dclservices/registration/
Version	1.0
Last update	2012-07-20

Contacts

Angel Genov	
Position	Affiliated researcher
Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia dcltools@dcl.bas.bg http://dcl.bas.bg/PersonalPages/angel/
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia est@dcl.bas.bg http://dcl.bas.bg
Availability start date	2012-06-01	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Web executable
Download location	http://dcl.bas.bg/dclservices/registration/
Execution location	http://dcl.bas.bg/dclservices/registration/

Metadata

Creation date	2012-07-20
----------------------	------------

Metadata last date updated	2013-01-31
-----------------------------------	------------

Validation

Validated	True
------------------	------

Usage

Foreseen use	NLP applications
Actual uses	NLP applications

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds Own funds
	Start date	2011-02-01
	End date	2013-01-30
Creation start date	2010-01-01	

Resource documentation

Documents	Manual	
	Title	Web-Based Infrastructure for Bulgarian Data Processing: User Guide
	URL	http://dcl.bas.bg/dclservices/WebInfrastructure-UserManual.pdf

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Media type	text
Output	Media type	text
Operating system	Linux	
Tool/service evaluation	Evaluated	True
	Level	Diagnostic
	Evaluators	Angel Genov

		Position	Assistant
		Contact	angel@dcl.bas.bg http://dcl.bas.bg/PersonalPages/angel/

8.17. TREFL – Translation Reference Library

General Information

Short name	TREFL
Description	TREFL is a portable, multifunctional database management application for Windows, having the combined characteristics of both a Translation Memory System (bilingual databases, fuzzy matching, concordance, alignment, importing and exporting translation memories, etc.) and those of an Internet/Desktop Search Engine (searching, like with Google search, all these words, this exact phrase, I'm feeling Lucky, etc.), plus some elements of semantic search. It is intended to be used as a simple, versatile, portable, effective and customizable reading, writing and translation aid tool capable of managing very large databases.
Identifier	817
Resource type	Tool/service
Tool/service type	Tool
URL	http://web.uni-plovdiv.bg/rousni/index_fr.htm
Version	1.0

Contacts

Roussi Nikolov	
Position	Associate Professor
Contact	Tsar Asen 24 4000 Plovdiv roussi.nikolov@gmail.com
Organization	Plovdiv University Paisii Hilendarski Department of Roman and Germanic Studies roussi.nikolov@gmail.com

Distribution

Availability	Available – restricted use	
IPR holder	Plovdiv University Paisii Hilendarski	
	Short name	PU
	Department name	Department of Roman and Germanic Studies
	Contact	24 Tsar Asen 4000 Plovdiv

		roussi.nikolov@gmail.com
--	--	--

Licences

CC-BY		
Restrictions of use	Academic – non-commercial use Inform licensor	
Access medium	Downloadable	
Download location	http://web.uni-plovdiv.bg/rousni/index.htm	
Execution location	http://web.uni-plovdiv.bg/rousni/index_fr.htm	
Signatories	Roussi Nikolov	
	Position	Associate Professor
	Contact	Tsar Asen 24 4000 Plovdiv roussi.nikolov@gmail.com
	Organization	Plovdiv University Paisii Hilendarski Department of Roman and Germanic Studies roussi.nikolov@gmail.com
	Plovdiv University Paisii Hilendarski	
	Contact	Tsar Asen 4000 Plovdiv roussi.nikolov@gmail.com

Metadata

Creation date	2013-01-29	
Metadata creators	Tsvetana Dimitrova	
	Position	Assistant Professor
	Contact	cvetana@dcl.bas.bg
	Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg
Metadata last date updated	2013-01-31	

Validation

Validated	True
------------------	------

Usage

--	--

Access tool	http://web.uni-plovdiv.bg/rousni/index_fr.htm
Foreseen use	Human use NLP applications
Actual uses	Human use
	NLP applications

Resource creation

Creation start date	2007-01-01
----------------------------	------------

Resource documentation

Reports	Nikolov, Roussi and Jean-Yves Dommergues. Les modules d'un système d'aide à la traduction en rapport avec la théorie interprétative. - Théorie, Littérature, Epistémologie, 25, 2008, pp. 105-123. Roussi Nikolov & Malina DITCHEVA, Една програма-помощник за превод, четене и писане, Plovdiv University "Paissii Hilendarski" - Bulgaria, Scientific Works, Vol. 45, Book 1, 2007 – Philology
Samples location	http://web.uni-plovdiv.bg/rousni/index_fr.htm
Tool documentation type	Online

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Bulgarian	
	Media type	text
	Modality type	Written language
	Language ID	bua en fr
	Segmentation level	Other Word
Output	UTF-8	
	Media type	text
Operating system	Windows	
Tool/service evaluation	Evaluated	True
	Level	Diagnostic Usage

8.18. SARP - Speech Analyzer Rapid Plot. Plotting vowels in F2-F1 scatter charts with multiple data sets

General Information

Short name	SARP
Description	The SaRP tool, which is an extension to the programme Speech Analyzer version 3 or later, allows managing databases of spoken language samples and creating informative charts in an easy and interactive manner. Key features: Computer generated feedback on vowel production by language learners. Designed for automatic or semi-automatic (interactive) retrieving of formant values. Easily creates, saves and opens vowel charts. Fully configurable and easy to use. Support for multiple data sets. Vowel charts comparison by superimposing control charts and user charts. Numerical or visual/graphical editing of the charts and quick-commands: create, move, delete, lock/unlock markers. Calculating and representing graphically the mean values. Integrated library of vocal samples.
Identifier	818
Resource type	Tool/service
Tool/service type	Tool
URL	http://web.uni-plovdiv.bg/rousni/sarp
Version	5.0.0

Contacts

Roussi Nikolov	
Position	Associate Professor
Contact	24 Tsar Asen 4000 Plovdiv roussi.nikolov@gmail.com
Organization	Plovdiv University Paisii Hilendarski Department of Roman and Germanic Studies roussi.nikolov@gmail.com

Distribution

Availability	Available – restricted use	
IPR holder	Plovdiv University Paisii Hilendarski	
	Short name	PU
	Department name	Department of Roman and Germanic Studies
	Contact	24 Tsar Asen 4000 Plovdiv roussi.nikolov@gmail.com

Licences

CC-BY	
Restrictions of use	Inform licensor
Access medium	Downloadable
Download location	http://web.uni-plovdiv.bg/rousni/sarp/download.html
Execution location	http://web.uni-plovdiv.bg/rousni/sarp/download.html
Signatories	Roussi Nikolov
	Position Associate Professor
	Contact 24 Tsar Asen 4000 Plovdiv roussi.nikolov@gmail.com
	Organization Plovdiv University Paisii Hilendarski Department of Roman and Germanic Studies roussi.nikolov@gmail.com
	Plovdiv University Paisii Hilendarski
	Short name PU
	Department name Department of Roman and Germanic Studies
	Contact 24 Tsar Asen 4000 Plovdiv roussi.nikolov@gmail.com
User nature	Academic

Metadata

Creation date	2013-01-29
Metadata last date updated	2013-01-31

Usage

Access tool	http://web.uni-plovdiv.bg/rousni/sarp/download.html
Foreseen use	Human use NLP applications
Actual uses	Human use
	NLP applications

Resource creation

Creation start date	2007-01-01
----------------------------	------------

Resource documentation

Reports	<p>Nikolov, R. & Dommergues & Élise RYST, SaRP: Un outil de représentations graphiques multi-points et multi-séries des formants vocaliques, Plovdiv University "Paissii Hilendarski" - Bulgaria, Scientific Works, Vol. 45, Book 1, 2007 – Philology</p> <p>Nikolov, R. & Nadine HERRY-BENIT, Spécificités méthodologiques de l'analyse des voyelles dans les voix de femmes, Plovdiv University "Paissii Hilendarski" - Bulgaria, Scientific Works, Vol. 46, Book 1, 2008 – Philology</p> <p>Nikolov, R. & Nadine HERRY-BENIT & Anne TORTEL, Positional determination of the quality of schwa in english, Plovdiv University "Paissii Hilendarski" - Bulgaria, Scientific Works, Vol. 47, Book 1, 2009 – Philology</p>
----------------	--

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Bulgarian	
	Media type	audio textNumerical
	Modality type	Spoken language
	Language ID	bg en fr
Output	Media type	image text textNumerical
	Modality type	Spoken language
Operating system	Windows	

8.19. RTComp - Real Time Comparison

General Information

Short name	RTComp
Description	RTComp allows effective management of multilingual databases of numerical speech models and graphical representations for direct visual comparison with the results of the real-time acoustic analysis of the language learners' speech.
Identifier	819
Resource type	Tool/service
Tool/service type	Tool
URL	http://web.uni-plovdiv.bg/rousni/rtcomp

Contacts

Roussi Nikolov	
Position	Associate Professor
Contact	Tsar Asen 24 4000 Plovdiv roussi.nikolov@gmail.com http://web.uni-plovdiv.bg/rousni/rtcomp
Organization	Plovdiv University Paisii Hilendarski Department of Roman and Germanic Studies roussi.nikolov@gmail.com

Distribution

Availability	Available – unrestricted use	
IPR holder	Plovdiv University Paisii Hilendarski	
	Short name	PU
	Department name	Department of Roman and Germanic Studies
	Contact	Tsar Asen 24 4000 Plovdiv roussi.nikolov@gmail.com http://web.uni-plovdiv.bg/rousni/rtcomp

Licences

CC-BY		
Restrictions of use	Academic – non-commercial use Inform licensor	
Access medium	Downloadable	
Download location	http://web.uni-plovdiv.bg/rousni/rtcomp	
Signatories	Roussi Nikolov	
	Position	Associate Professor
	Contact	24 Tsar Asen 4000 Plovdiv roussi.nikolov@gmail.com
	Plovdiv University Paisii Hilendarski	
	Short name	PU
	Department name	Department of Roman and Germanic Studies

	Contact	24 Tsar Asen 4000 Plovdiv roussi.nikolov@gmail.com
--	----------------	--

Metadata

Creation date	2013-01-29
Metadata last date updated	2013-01-31

Usage

Access tool	http://web.uni-plovdiv.bg/rousni/rtcomp/
Foreseen use	Human use NLP applications
Actual uses	Human use NLP applications

Resource creation

Creation start date	2012-01-01
----------------------------	------------

Resource documentation

Samples location	http://web.uni-plovdiv.bg/rousni/rtcomp/
Tool documentation type	Online Other

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Bulgarian	
	Media type	audio
	Modality type	Spoken language
	Language ID	bg en fr

8.20. Corpus of Spoken Bulgarian

General Information

Short name	SpokenBg
Description	The Corpus of Spoken Bulgarian (SpokenBg) is a selection of data of spoken Bulgarian language incl. data from interviews, media and formal speech, student speech, academic speech, colloquial speech. The total size of the corpus is 523,128 signs as of the end of 2012. Part of it contains edited versions of transcripts of conversational speech being converted from a semi-phonetic transcription to standard orthography with original semi-phonetic transcripts presented together with the edited versions in a paragraph-aligned display.
Identifier	820
Resource type	Corpus
URL	http://bgspeech.net/
Version	4.0

Contacts

Yovka Tisheva	
Position	Associate Professor
Contact	15 Tsar Osvoboditel Blvd. 1504 Sofia yovka.tisheva@gmail.com http://bgspeech.net/
Organization	Sofia University "St. Kliment Ohridski" Department for Bulgarian Language yovka.tisheva@gmail.com

Distribution

Availability	Available – unrestricted use	
IPR holder	Sofia University "St. Kliment Ohridski"	
	Short name	SU
	Department name	Department for Bulgarian Language
	Contact	15 Tsar Osvoboditel Blvd. 1504 Sofia yovka.tisheva@gmail.com http://bgspeech.net/
Availability start date	2013-01-21	

Licences

CC-BY-NC-SA	
Access medium	Accessible through interface

Metadata

Creation date	2013-01-27
Metadata last date updated	2013-02-01

Validation

Validated	True
------------------	------

Resource creation

Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30
	Models and tools for spoken communication of contemporary Bulgarian language	
	Funding type	National funds
	Country	Bulgaria
Creation start date	2010-01-01	

Resource documentation

Reports	<p>Atanasov, Atanas. Encoding Bulgarian Colloquial Speech Using TEI Specification. - In: Computer Applications in Slavic Studies. "Boyan Penev" Publishing Center, Sofia, 2006, pp. 233-240.</p> <p>Атанасов, Атанас. Проблеми при създаването на езикови корпуси с транскрибирана българска разговорна реч. - В: Паисиеви четения. Научни трудове. Том 44, кн. 1, сб. А, 2006. УИ "Паисий Хилендарски", Пловдив, 2006, с. 289-296.</p> <p>Тишева, Йовка, Марина Джонова. Електронни ресурси за българската разговорна реч (инициативата BgSpeech). - Littera et Lingua, лято 2010.</p>
----------------	---

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
Modality	Modality type	Spoken language

	Modality type	Spoken language
Size	523,128 phonetic units	
Character encoding	UTF-8	
Annotation	Speech annotation – speaker turns	
	Segmentation level	Paragraph Sentence Word
Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
Size	10 hours	

8.21. Corpus of Colloquial Bulgarian

General Information

Short name	BgSpeech
Description	The Corpus of Colloquial Bulgarian is a selection of data of oral forms of contemporary Bulgarian language amounting to 357,584 signs.
Identifier	821
Resource type	Corpus
URL	http://www.bgspeech.net/index.html
Version	4.0
Last update	2013-01-15

Contacts

Yovka Tisheva	
Position	Associate Professor
Contact	15 Tsar Osvoboditel Blvd. 1504 Sofia yovka.tisheva@gmail.com
Organization	Sofia University St. Kliment Ohridski Department of Bulgarian Language yovka.tisheva@gmail.com

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

IPR holder	Sofia University St. Kliment Ohridski	
	Short name	SU
	Department name	Department of Bulgarian Language
	Contact	15 Tsar Osvoboditel Blvd. 1504 Sofia yovka.tisheva@gmail.com
Availability start date	2013-01-21	

Licences

CC-BY-NC-SA	
Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Download location	http://www.bgspeech.net/bg/resources.html
Execution location	http://www.bgspeech.net/bg/resources.html

Metadata

Creation date	2013-01-29	
Metadata creators	Tsvetana Dimitrova	
	Position	Assistant Professor
	Contact	52 Shipchenski prohod Blvd. 1113 Sofia cvetana@dcl.bas.bg
	Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg
Metadata last date updated	2013-02-01	

Validation

Validated	True
------------------	------

Usage

Access tool	http://www.bgspeech.net/bg/resources.html
Foreseen use	Human use
Actual uses	Human use

Resource creation

Resource creator	Sofia University St. Kliment Ohridski	
	Short name	SU
	Department name	Department of Bulgarian Language
	Contact	15 Tsar Osvoboditel Blvd. 1504 Sofia yovka.tisheva@gmail.com
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30
	Syntactic characteristics of the contemporary spoken Bulgarian	
	Funding type	National funds
	Country	Bulgaria
	Maintaining and updating the data base of contemporary Bulgarian language	
	Funding type	National funds
	Country	Bulgaria
	Annotating corpora of spoken Bulgarian	
	Funding type	National funds
	Country	Bulgaria
	Elaboration of the transcription system for the spoken Bulgarian	
	Funding type	National funds
	Country	Bulgaria
Creation start date	2000-05-01	

Resource documentation

Reports	<p>Atanas Atanasov. Encoding Bulgarian Colloquial Speech Using TEI Specification. Computer Applications in Slavic Studies. "Boyan Penev" Publishing Center, Sofia, 2006, pp. 233-240</p> <p>Атанас Атанасов. Проблеми при създаването на езикови корпуси с транскрибирана българска разговорна реч. Паисиеви четения. Научни трудове, том 44, кн. 1, сб. А, 2006. УИ "Паисий Хилендарски", Пловдив, 2006, 289-296</p> <p>Йовка Тишева, Марина Джонова. Електронни ресурси за българската разговорна реч (инициативата BgSpeech). Littera et Lingua, лято 2010.</p>
----------------	--

	Йовка Тишева, Марина Джонова. Корпус с устна българска реч – специфика и структура. Български език 58 (2011), 3, 34-53
Samples location	http://www.bgspeech.net/bg/resources.html
Tool documentation type	Online

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
	Language script	Cyrillic
Modality	Modality type	Spoken language
Size	357,584 phonetic units	
Character encoding	UTF-8	
Annotation	Speech annotation – speaker turns	
	Segmentation level	Paragraph Phoneme Prosodic boundaries Sentence Word

8.22. Dictionary of Synonyms in Bulgarian Language

General Information

Short name	DSBL
Description	The Dictionary of Synonyms in Bulgarian Language covers the body of synonyms in Modern Bulgarian. It contains ca. 27,000 unique word forms pertaining to four parts-of-speech, distributed into synonym sets, as follows: verbs: 2,137 synonym sets, containing a total number of 10,000 words. nouns: 3 240 synonym sets, containing a total number of over 12,000 words. adjectives: 2 496 synonym sets, containing a total number of over 10,000 words. adverbs: 910 synonym sets, containing a total number of over 3,800 words. The words are given in their basic form. Verbs are given only in their imperfective form while perfectives are marked by a grammatical note in parentheses after the verb form, as in: повеселя се (само св.). One word may be a member of more than one synonym set depending on its meaning. When searching in the database for a certain word, synonym sets, containing the searched word, are displayed in a row in ascending order. To distinguish between different meanings of a polysemous word that is a member of more than one synonym set, short explanations, grammatical or stylistic notes are placed in parentheses after the word, if necessary. Homonymy is marked by a superscript.
Identifier	822

Resource type	Lexical conceptual resource
Lexical conceptual resource type	Lexicon
URL	http://infolex.ibl.bas.bg/synomthes/home.seam?cid=18
Version	1.0
Last update	2013-01-10

Contacts

Diana Blagoeva	
Position	Associate Professor
Contact	52 Shipchenski Prohod Blvd., Bl. 17 1113 Sofia d.blagoeva@ibl.bas.bg http://www.ibl.bas.bg/en/people_en2_cv_Blagoeva.htm
Organization	Institute for Bulgarian Language Department of computational linguistics, Department of Bulgarian lexicology and lexicography ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of computational linguistics, Department of Bulgarian lexicology and lexicography
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg
Availability start date	2013-01-21	

Licences

CC-BY-NC-ND	
Restrictions of use	Academic – non-commercial use No redistribution
Access medium	Accessible through interface
Download location	http://infolex.ibl.bas.bg/synomthes/home.seam?cid=18

Metadata

Creation date	2013-01-27	
Metadata creators	Tsvetana Dimitrova	
	Position	Assistant Professor
	Contact	cvetana@dcl.bas.bg
	Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg
Metadata last date updated	2013-01-31	

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use
Actual uses	Human use

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Bulgarian Lexicology and Lexicography
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia d.blagoeva@ibl.bas.bg http://ibl.bas.bg/en/index.htm
Funding projects	CESAR: Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30
	Complex Study of Bulgarian Lexis and Phraseology as Part of the Cultural Heritage and National Identity	
	Funding type	National funds

	Country	Bulgaria
Creation start date	2009-12-17	

Resource documentation

Reports	Blagoeva, Diana, Sia Kolkovska. „Инфолекс” – информационен лингвистичен портал за целите на българската лексикология, фразеология и лексикография. – Littera et Lingua Electronic Journal of Humanities, 2012, Fall, 10 p. ISSN: 1312-6172 http://www.slav.uni-sofia.bg/lilijournal
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Lexicon	
Lexical conceptual resource encoding	Encoding level	Semantics
	Linguistic information	Semantics – relations – synonyms

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
	Language script	Cyrillic
	Size	27,000 words
Modality	Modality type	Written language
Size	27,000 words	
Character encoding	UTF-8	

8.23. Dictionary of Antonyms in Bulgarian Language

General Information

Short name	DABL
Description	The Dictionary of Antonyms in Bulgarian Language covers the body of synonyms in Modern Bulgarian. It contains about 8,500 unique word forms pertaining to four parts-of-speech, distributed into 3,644 antonym sets, as follows: verbs: 571 antonym sets and a total number of over 3,000 words; nouns: 1,399 antonym sets and a total number of over 5,000 words; adjectives: 1,092 antonym sets and a total number of over 4,100 words; adverbs: 582 antonym sets and a total number of over 2,100 words. The words are given in their basic form. Verbs are given only in imperfective form, while perfectives are marked by a grammatical note in parentheses after the verb, as in: <i>дойда (само св.)</i> . One word may be a member of more than one antonym set depending on its meaning. When searching in the

	database for a certain word, antonym sets, containing the searched word, are displayed in a row in ascending order. To specify the stylistic or grammatical characteristics of words in an antonym set, if necessary, notes are placed in parentheses after the word.
Identifier	823
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Lexicon
URL	http://infolex.ibl.bas.bg/synonthes/antonimSearchPage.seam?cid=19
Version	1.0
Last update	2013-01-10

Contacts

Diana Blagoeva	
Position	Associate Professor
Contact	52 Shipchenski Prohod Blvd., Bl. 17 1113 Sofia d.blagoeva@ibl.bas.bg http://www.ibl.bas.bg/en/people_en2_cv_Blagoeva.htm
Organization	Institute for Bulgarian Language Department of computational linguistics, Department of Bulgarian lexicology and lexicography ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Bulgarian Lexicology and Lexicography
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg
Availability start date	2013-01-21	

Licences

CC-BY-NC-ND	
Restrictions of use	Academic – non-commercial use

	No redistribution
Access medium	Accessible through interface
Download location	http://infolex.ibl.bas.bg/synomthes/antonimSearchPage.seam?cid=19

Metadata

Creation date	2013-01-27
Metadata last date updated	2013-01-31

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use
Actual uses	Human use

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics Department of Bulgarian Lexicology and Lexicography
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia d.blagoeva@ibl.bas.bg http://ibl.bas.bg/en/index.htm
Funding projects	Complex Study of Bulgarian Lexis and Phraseology as a Part of the Cultural Heritage and National Identity	
	Funding type	National funds
	Country	Bulgaria
	CESAR: Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30
Creation start date	2009-12-17	

Resource documentation

Reports	Blagoeva, Diana, Sia Kolkovska. „Инфолекс” – информационен лингвистичен портал за целите на българската лексикология, фразеология и лексикография. – Littera et Lingua Electronic Journal of Humanities, 2012, Fall, 10 p. ISSN: 1312-6172 http://www.slav.uni-sofia.bg/lilijournal
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Lexicon	
Lexical conceptual resource encoding	Encoding level	Semantics
	Linguistic information	Semantics – relations – antonyms

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
	Language script	Cyrillic
Modality	Modality type	Written language
Size	3,644 words	
Character encoding	UTF-8	

8.24. Register of Phraseologisms in Bulgarian Language

General Information

Short name	RPhBL
Description	The Register of Phraseologisms in Bulgarian Language covers the body of phraseologisms in Bulgarian language. It contains phraseological units, as defined by and found in the main Bulgarian phraseological dictionaries. The total number of the phraseological units is over 10,000, and the unique words – over 6,400. When searching in the register for a word, a list of all phraseological units containing the searched word (or a part of it) is composed. Each phraseological unit has a reference note to the source from which it was extracted. Thus, the register provides information on lexicographic editions, in which the meaning of the phraseological unit can be found.
Identifier	824
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Lexicon

URL	http://infolex.ibl.bas.bg/PhrasThes/searchPhrasesPage.seam
Version	1.0
Last update	2013-01-10

Contacts

Diana Blagoeva	
Position	Associate Professor
Contact	52 Shipchenski Prohod Blvd., Bl. 17 1113 Sofia d.blagoeva@ibl.bas.bg http://www.ibl.bas.bg/en/people_en2_cv_Blagoeva.htm
Organization	Institute for Bulgarian Language Department of computational linguistics, Department of Bulgarian lexicology and lexicography ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Bulgarian Lexicology and Lexicography
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg
Availability start date	2013-01-21	

Licences

CC-BY-NC-ND	
Restrictions of use	Academic – non-commercial use No redistribution
Access medium	Accessible through interface
Download location	http://infolex.ibl.bas.bg/PhrasThes/searchPhrasesPage.seam

Metadata

Creation date	2013-01-27

Metadata creators	Tsvetana Dimitrova	
	Position	Associate Professor
	Contact	cvetana@dcl.bas.bg
	Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg
Metadata last date updated	2013-01-31	

Validation

Validated	True
------------------	------

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Lexicology and Lexicography
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia d.blagoeva@ibl.bas.bg http://ibl.bas.bg/en/index.htm
Funding projects	Complex Study of Bulgarian Lexis and Phraseology as Part of the Cultural Heritage and National Identity	
	Funding type	National funds
	Country	Bulgaria
	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30
Creation start date	2009-12-17	

Resource documentation

Reports	Blagoeva, Diana, Sia Kolkovska. „Инфолекс” – информационен лингвистичен портал за целите на българската лексикология, фразеология и лексикография. – Littera et Lingua Electronic Journal of Humanities, 2012, Fall, 10 p. ISSN: 1312-6172 http://www.slav.uni-sofia.bg/lilijournal
----------------	---

Lexical conceptual resource

Lexical conceptual resource type	Lexicon	
Lexical conceptual resource encoding	Encoding level	Semantics

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
Modality	Modality type	Written language
Size	10,000 entries	
Character encoding	UTF-8	

8.25. Dictionary of Neologisms in Bulgarian Language

General Information

Short name	DNBL
Description	The Dictionary of Neologisms in Bulgarian Language contains over 2,200 new words and 160 new multiword units (compounds and terminological units) that have entered the Bulgarian language in the past 20 years. Each entry contains information about: part-of-speech (for lexemes); origin (for borrowed words); stylistic and grammatical notes; lexical meaning of the unit; synonyms and antonyms (if available). If necessary, short examples (phrases or sentences) are given to illustrate the use of the neologism in context.
Identifier	825
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Lexicon
URL	http://infolex.ibl.bas.bg/PhrasThes/searchNeologPage.seam?cid=17
Version	1.0
Last update	2013-01-10

Contacts

Diana Blagoeva	
Position	Associate Professor
Contact	52 Shipchenski Prohod Blvd., Bl. 17 1113 Sofia

	d.blagoeva@ibl.bas.bg http://www.ibl.bas.bg/en/people_en2_cv_Blagoeva.htm
Organization	Institute for Bulgarian Language Department of computational linguistics, Department of Bulgarian lexicology and lexicography ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Bulgarian Lexicology and Lexicography
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia d.blagoeva@ibl.bas.bg
Availability start date	2013-01-21	

Licences

CC-BY-NC-ND	
Restrictions of use	Academic – non-commercial use No redistribution
Access medium	Accessible through interface
Download location	http://infolex.ibl.bas.bg/PhrasThes/searchPhrasesPage.seam

Metadata

Creation date	2013-01-27	
Metadata creators	Tsvetana Dimitrova	
	Position	Associate Professor
	Contact	cvetana@dcl.bas.bg
	Organization	Institute for Bulgarian Language Department of Computational Linguistics cvetana@dcl.bas.bg
Metadata last date updated	2013-02-01	

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use
Actual uses	Human use

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia d.blagoeva@ibl.bas.bg http://ibl.bas.bg/en/index.htm
Funding projects	Complex Study of Bulgarian Lexis and Phraseology as Part of the Cultural Heritage and National Identity	
	Funding type	National funds
	Country	Bulgaria
	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30
Creation start date	2009-12-17	

Resource documentation

Reports	Blagoeva, Diana, Sia Kolkovska. „Инфолекс” – информационен лингвистичен портал за целите на българската лексикология, фразеология и лексикография. – Littera et Lingua Electronic Journal of Humanities, 2012, Fall, 10 p. ISSN: 1312-6172 http://www.slav.uni-sofia.bg/lilijournal
----------------	---

Lexical conceptual resource

Lexical conceptual resource type	Lexicon	
Lexical conceptual resource encoding	Encoding level	Semantics

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
	Language script	Cyrillic
Modality	Modality type	Written language
Size	2,360 entries	
Character encoding	UTF-8	

8.26. Bulgarian Spell Checker for Mac OS

General Information

Short name	MacEst
Description	<p>The system for automatic spelling checking MacEst for Mac OS X detects and marks the incorrectly written words in a text and suggests the most probable candidates to correct the errors. MacEst offers the entire potential of a contemporary spelling correction: proficiently compiled dictionary, which contains over a million and a half words, and replacement suggestions ordered according to their probability. MacEst is based on the Electronic Grammar Dictionary of Bulgarian, developed at the Department of Computational Linguistics at the Institute for Bulgarian Language, which contains over 85,000 words. It contains logic for detection of careless mistakes (wrong key pressed, letter swapping, skipped or extra letters), identifies errors of ignorance, and integrates perfectly into the dictionaries used in Mac OS. MacEst uses an extremely fast and effective method for searching and detecting the correct words regardless of the text size. The functionality of the product is realized through the use of minimal acyclic deterministic automata and Levenshtein automata, which allow maximum speed, precision and coverage. Advantages: MacEst offers the entire potential of the contemporary spelling checking and correction. Together with the proficiently compiled dictionary the product is capable of finding replacement suggestions, which are ranked by probability. Representativeness: covers the basic wordstock of Bulgarian. Precision: all words are checked by experts. Convenience: the replacement candidates are ranked by probability. Interface: NSSpellServer. MacEst is available for all applications that use Mac OS X spell checking system.</p>
Identifier	826
Resource type	Tool/service
Tool/service type	Tool
URL	http://dcl.bas.bg/en/MacEst-en.html
Version	2.0

Contacts

Svetla Koeva

Position	Professor
Contact	52 Shipchenski prohod Blvd. 1113 Sofia svetla@dcl.bas.bg http://dcl.bas.bg/PersonalPages/svetla/svetla.html
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

CC-BY-NC	
Restrictions of use	Other
Access medium	Downloadable
Download location	http://dcl.bas.bg/sites/default/files/webfm/MacEst/MacEst1-1.0-beta1.dmg

Metadata

Creation date	2013-01-29	
Metadata creators	Tsvetana Dimitrova	
	Position	Assistant Professor
	Contact	52 Shipchenski prohod Blvd. 1113 Sofia cvetana@dcl.bas.bg
	Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg
Metadata last date updated	2013-02-01	

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use
Actual uses	Human use

Resource creation

Funding projects	Web applications for editing Bulgarian texts	
	Funding type	National funds
	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds

Resource documentation

Reports	Oliva, Karel, Svetla Koeva. Sintaksis na nevazmojnoto (Syntax of the Impossible). - Balgarski ezik, 2009, 3, pp. 7-17. ISSN 0005-4283. (In Bulgarian)
Tool documentation type	Online

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Bulgarian	
	Media type	text
	Language ID	bg
Output	Media type	text
Operating system	Mac OS	
Tool/service evaluation	Evaluated	True
	Level	Diagnostic
	Evaluators	Angel Genov
	Contact	angel.genov@gmail.com

8.27. Wiki1000+ corpus with annotated MWEs

General Information

Short name	Wiki1000+
Description	Wiki1000+ is a corpus of articles from Wikipedia, compiled for the purposes of the study of multiword expressions (MWEs) in Bulgarian. The Wiki1000+ corpus contains 6,311 text samples with at least 1,000 tokens each, amounting to 13.4 million tokens. The corpus is a part of the Bulgarian National Corpus. Wiki1000+ is annotated with the following linguistic information: sentence boundaries, tokenisation, lemmatisation, POS tagging, and

	MWE annotation. MWE annotation includes MWE id, labelling the components of the MWE and determining the type of the MWE according to a classification based on idiomaticity.
Identifier	827
Resource type	Corpus
URL	http://dcl.bas.bg/en/wikiCorpus_en.html

Contacts

Svetla Koeva	
Position	Professor
Contact	52 Shipchenski prohod Blvd. 1113 Sofia svetla@dcl.bas.bg http://dcl.bas.bg/PersonalPages/svetla/svetla.html

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

CC-BY	
Restrictions of use	Academic – non-commercial use Attribution
Access medium	Downloadable
Download location	http://dcl.bas.bg/BulNC-registration/feeds/CESAR/Wiki1000.zip

Metadata

Creation date	2013-01-30
Metadata last date updated	2013-02-01

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd. 1113 Sofia ibl@ibl.bas.bg

		http://ibl.bas.bg/en/index.htm
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30
	Bulgarian National Corpus project	
	Project short name	BulNC
	Funding type	National funds

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
	Language script	Cyrillic
Modality	Modality type	Written language
Size	13,400,400 tokens	
Annotation	Segmentation	
	Segmentation level	Sentence Word
	Lemmatization	
	Morphosyntactic annotation – POS tagging	

8.28. The Bulgarian-English Sentence- and Clause-Aligned Corpus

General Information

Short name	BulEnAC
Description	The Bulgarian-English Sentence- and Clause-Aligned Corpus (BulEnAC) is an excerpt from the Bulgarian-English Parallel Corpus – a part of the Bulgarian National Corpus (BulNC). The Bulgarian-English Parallel Corpus has been processed at several levels: tokenisation, sentence splitting, lemmatisation. The processing has been performed using the Bulgarian language processing chain for the Bulgarian part and Apache OpenNLP and Stanford CoreNLP for the English part. The BulEnAC consists of 176,397 tokens for Bulgarian and 190,468 for English (366,865 tokens altogether). Sentences are 30,385

	(14,667 Bulgarian sentences (12.02 words per sentence on average) and 15,718 English sentences (12.11 words per sentence). The average number of clauses in a sentence in the Bulgarian part is 1.67 compared to 1.85 clauses per sentence for the English part. The texts are distributed over five broad categories, called 'styles': administrative, fiction, science, journalism, and subtitles. The corpus is represented in XML format and is supplied with various linguistic annotation – monolingual for both Bulgarian and English (sentence splitting, tokenisation, lemmatisation, POS and grammatical tagging), and parallel (sentence and clause alignment).
Identifier	828
Resource type	Corpus
URL	http://dcl.bas.bg/en/ClauseAlignedCorpus_en.html
Version	1.0

Contacts

Svetla Koeva	
Position	Professor
Contact	52 Shipchenski prohod Blvd. 1113 Sofia svetla@dcl.bas.bg http://dcl.bas.bg/PersonalPages/svetla/svetla.html

Distribution

Availability	Available – restricted use
---------------------	----------------------------

Licences

Restrictions of use	Other
Access medium	Downloadable
Download location	http://dcl.bas.bg/BulNC-registration/feeds/CESAR/BulEnAC.zip
Execution location	http://dcl.bas.bg/BulNC-registration/feeds/CESAR/BulEnAC.zip
User nature	Academic

Metadata

Creation date	2013-01-30
Metadata last date updated	2013-02-01

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL

	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd. 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/en/index.htm
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30
	Bulgarian National Corpus project	
	Project short name	BulNC
	Funding type	National funds

Resource documentation

Reports	Koeva, Svetla, Borislav Rizov, Ekaterina Tarpomanova, Tsvetana Dimitrova, Rositsa Dekova, Ivelina Stoyanova, Svetlozara Leseva, Hristina Kukova, Angel Genov. Bulgarian-English Sentence- and Clause-Aligned Corpus. – In: Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2), Lisbon: Edicoes Colibri, 2012, pp. 51-62.
Tool documentation type	Online

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Languages	Bulgarian	
	Language ID	bg
	Language script	Cyrl
	Size	176,397 tokens
	English	
	Language ID	en
	Language	Latn

	script	
	Size	190,468 tokens
Modality	Modality type	Written language
Size	30,385 sentences	
Annotation	Segmentation	
	Segmentation level	Clause Paragraph Sentence Word
	Alignment	
	Segmentation level	Clause Sentence
	Semantic annotation – word senses	
	Segmentation level	Word
	Morphosyntactic annotation – POS tagging	
	Segmentation level	Word

8.29. Multilingual dictionaries

General Information

Description	The set of multilingual dictionaries covers all pairs of languages among the following: Bulgarian, English, German, Romanian, Greek, and Polish. The main source of the dictionaries is Wikipedia – translations of article titles and category labels. The dictionaries include single words, MWEs and phrases but are predominantly phrase-to-phrase. The following sets of dictionaries are included in the pack: General bilingual dictionaries for each pair of languages; Bilingual dictionaries of personal names for each pair of languages; Bilingual dictionaries of organisations for each pair of languages; Bilingual dictionaries of toponyms for each pair of languages. The dictionaries are stored in plain text format for easy and flexible storage and processing.
Identifier	829
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Lexicon
URL	http://dcl.bas.bg/en/multilingualDictionary_en.html
Version	1.0

Contacts

Svetla Koeva	
---------------------	--

Position	Professor
Contact	52 Shipchenski prohod Blvd. 1113 Sofia svetla@dcl.bas.bg http://dcl.bas.bg/PersonalPages/svetla/svetla.html

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/en/index.htm

Licences

CC-BY-NC	
Restrictions of use	Academic – non-commercial use
Access medium	Downloadable
Download location	http://dcl.bas.bg/BulNC-registration/feeds/CESAR/MultilingualDictionary.zip

Metadata

Creation date	2013-01-30
Metadata last date updated	2013-01-31

Resource creation

Resource creator	Tetracom Interactive Solutions Ltd.	
	Short name	Tetracom
	Contact	18 Prof. Nikolay Gentchev Str. 1700 Sofia info@tetracom.com http://www.tetracom.com
Funding projects	Central and South-East European Resources	
	Project short name	CESAR

	URL	http://cesar.nytud.nu
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30

Lexical conceptual resource

Lexical conceptual resource type	Lexicon	
Lexical conceptual resource encoding	Encoding level	Morphology Other Semantics
	Linguistic information	Lemma – compounds Lemma – multi word units Part of speech Semantics – semantic class

Texts

Media type	text	
Linguality type	Multilingual	
Multilinguality type	Parallel	
Languages	Bulgarian	
	Language ID	bg
	Language script	Cyrillic
	English	
	Language ID	en
	Language script	Latin
	German	
	Language ID	de
	Language script	Latin
	Romanian	
	Language ID	ro
	Language script	Latin
	Greek, Modern (1453-)	
	Language ID	el

	Polish	
	Language ID	pl
	Language script	Latin
Modality	Modality type	Written language
Size	173 mb	
Text format	text	

8.30. Bulgarian MWE dictionary

General Information

Description	The Bulgarian dictionary of MWEs includes 27,744 MWEs altogether which are divided into 13 categories based on their idiomatity which is evaluated with respect to the following features: whether the MWE is a named entity; whether the MWE contains a reference to a named entity; the degree to which the meaning of the MWE is compositional and transparent. The MWEs are extracted from several sources: Wikipedia, the Thesaurus of Bulgarian (1994) and other printed dictionaries and electronic corpora. The MWEs are manually verified and classified into categories.
Identifier	830
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Lexicon
URL	http://dcl.bas.bg/en/mweDictionary_en.html
Version	1.0

Contacts

Ivelina Stoyanova	
Position	Affiliated researcher
Contact	52 Shipchenski prohod Blvd. 1113 Sofia iva@dcl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics

	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/en/index.htm
--	----------------	--

Licences

CC-BY-NC	
Restrictions of use	Academic – non-commercial use
Download location	http://dcl.bas.bg/BulNC-registration/feeds/CESAR/MWEDicts.zip

Metadata

Creation date	2013-01-30
Metadata last date updated	2013-02-01

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use
	NLP applications

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd. 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/en/index.htm
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu
	Funding type	EU funds

	Start date	2011-02-01
	End date	2013-01-30

Lexical conceptual resource

Lexical conceptual resource type	Lexicon	
Lexical conceptual resource encoding	Encoding level	Other Semantics
	Linguistic information	Lemma – multi word units Semantics – semantic class

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
Size	27,744 multi-word units	

8.31. bgMWE – tool for MWE recognition

General Information

Short name	bgMWE
Description	bgMWE is a tool for corpus processing and MWE recognition and tagging. It is developed in Java and is thus platform independent. bgMWE comprises a set of modules which can be applied for particular NLP tasks. It is largely language independent and can work either in resource-light mode, or its performance can be boosted by employing lexical resources. The system includes the following modules: Web crawler for Wikipedia; Extraction of lexical data – lists of words and MWEs; Converter between formats – vertical format, XML, etc.; Preprocessing module – applying a chunker, a tagger, etc.; Collection of frequency data; MWE recognition and tagging; Further improvement of bgMWE is planned in the following directions: improving efficiency; implementing various methods for MWE recognition; developing a visualisation module or integrating existing open source visualisation methods; module for extensive evaluation.
Identifier	831
Resource type	Tool/service
Tool/service type	Tool
URL	http://dcl.bas.bg/en/bgMWE_en.html
Version	1.0

Contacts

Ivelina Stoyanova	
Position	Affiliated researcher
Contact	iva@dcl.bas.bg http://dcl.bas.bg/en/people_en/iva.html
Organization	Institute for Bulgarian Language Department of Computational Linguistics iva@dcl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Department name	Department of computational linguistics
	Contact	dcl@dcl.bas.bg http://dcl.bas.bg

Licences

CC-BY-NC	
Restrictions of use	Academic – non-commercial use
	Other
Access medium	Downloadable
Download location	http://dcl.bas.bg/BulNC-registration/feeds/CESAR/bgMWE.zip

Metadata

Creation date	2013-01-30
Metadata last date updated	2013-02-01

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use
	NLP applications
Actual uses	Human use
	NLP applications

Resource creation

--	--

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd. 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/en/index.htm
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu/
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30

Tool/service

Tool/service type	Tool	
Language dependent	True	
Input	Bulgarian	
	Media type	text
	Modality type	Written language
	Language ID	bg
Output	Bulgarian	
	Media type	text
	Language ID	bg
Tool/service creation	Implementation language	Java

8.32. TextMatch

General Information

Description	TextMatch is a web service that combines language independent ways of computing the similarity between two documents using powerful linguistic tools, and provides you different measures of the document similarity. It returns a percentage reflecting the probability that one document is similar to the other. TextMatch can compare documents in different formats (such as Microsoft document formats, OpenDocument Format, Portable Document Format, Electronic Publication Format, HyperText Markup Language, Rich Text Format,
--------------------	--

	Text formats).TextMatch recognizes the language of the document using two-stage language detection system. Specific language tokenizers, lemmatizers and other analyzers are utilized for English, Bulgarian, German, French, and Russian. A language independent comparison algorithm is used if one of the uploaded documents is in another language.
Identifier	832
Resource type	Tool/service
Tool/service type	Service
URL	http://www.textmatch.eu/
Version	1.0
Last update	2013-01-30

Contacts

Svetla Koeva	
Position	Professor
Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia svetla@dcl.bas.bg http://dcl.bas.bg/PersonalPages/svetla/svetla.html
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/en/index.htm

Licences

Restrictions of use	Other
Access medium	Accessible through interface
	Other

Metadata

Creation date	2013-01-31	
Metadata creators	Tsvetana Dimitrova	
	Position	Assistant Professor
	Contact	cvetana@dcl.bas.bg
	Organization	Institute for Bulgarian Language Department of Computational Linguistics cvetana@dcl.bas.bg
Metadata last date updated	2013-01-31	

Usage

Foreseen use	Human use
Actual uses	Human use

Resource creation

Resource creator	Tetracom Interactive Solutions	
	Short name	Tetracom
	Contact	18 Prof. Nikolay Gentchev Str. 1700 Sofia info@tetracom.com http://www.tetracom.com/

Tool/service

Tool/service type	Service	
Language dependent	False	
Input	Bulgarian	
	Media type	text
	Modality type	Written language
	Language ID	bg en fr de ru
Output	Media type	textNumerical
	Modality type	Other

8.33. Bulgarian Grammar checker web service

General Information

Short name	WinEst+
Description	The Bulgarian Grammar checker is based on a language model derived from the frequency list of the annotated Bulgarian National Corpus. It checks 893,626,788 3-grams with POS tags, including punctuation. The results show the probability of an arbitrary 3-gram with part-of-speech tags to be valid in the language model. The language model is executed in the form of finite automata. For each sentence, the model consecutively applies 3-grams, and those that are below the threshold are flagged as potential errors.
Identifier	833
Resource type	Tool/service
Tool/service type	Service
URL	http://dcl.bas.bg/est/grammarcheck.php
Version	1.0
Last update	2013-01-30

Contacts

Angel Genov	
Position	Affiliated researcher
Contact	angel@dcl.bas.bg http://dcl.bas.bg/PersonalPages/angel/
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg

Licences

CC-BY-NC	
Restrictions of use	Academic – non-commercial use Other

Access medium	Accessible through interface
Execution location	http://dcl.bas.bg/est/grammarcheck.php

Metadata

Creation date	2013-01-31	
Metadata creators	Tsvetana Dimitrova	
	Position	Assistant Professor
	Contact	cvetana@dcl.bas.bg
	Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg
Metadata last date updated	2013-01-31	

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use
	NLP applications

Resource creation

Resource creator	Angel Genov	
	Position	Affiliated researcher
	Contact	angel@dcl.bas.bg
	Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg
	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/en/index.htm
	Funding projects	
	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu

	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30
	Web applications for editing Bulgarian texts	
	URL	http://dcl.bas.bg/est/index_en.php
	Funding type	National funds

Resource documentation

Reports	Oliva, Karel, Svetla Koeva. Sintaksis na nevazmojnoto (Syntax of the Impossible). - Balgarski ezik, 2009, 3, pp. 7-17. ISSN 0005-4283. (In Bulgarian)
----------------	---

Tool/service

Tool/service type	Service	
Language dependent	True	
Input	Bulgarian	
	Media type	text
	Modality type	Written language
	Language ID	bg
Output	Bulgarian	
	Media type	text
	Modality type	Written language
	Language ID	bg

8.34. N-grams from Bulgarian National Corpus

General Information

Short name	BgNgrams
Description	BgNgrams lists are extracted from the current version of the Bulgarian National Corpus (with a core Bulgarian part containing over 1.2 billion words). The n-grams involves both lemmas (n-gram lemma) and word forms (n-gram word form). n-grams can be 1-grams, 2-grams, 3-grams, 4-grams, 5-grams. The n-gram language models (1-5) are in the standard ARPA text and binary format.
Identifier	834
Resource type	Corpus
URL	http://dcl.bas.bg/Resources/NGrams/
Version	1.0

Last update	2013-01-30
--------------------	------------

Contacts

Svetla Koeva	
Position	Professor
Contact	svetla@dcl.bas.bg http://dcl.bas.bg/PersonalPages/svetla/svetla.html
Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/en/index.htm

Licences

CC-BY-NC	
Restrictions of use	Academic – non-commercial use Other
Access medium	Downloadable
Download location	http://dcl.bas.bg/Resources/NGrams/

Metadata

Creation date	2013-01-31
Metadata last date updated	2013-01-31

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use

	NLP applications
--	------------------

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Computational Linguistics
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/en/index.htm
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu
	Funding type	EU funds
	Start date	2011-02-01
	End date	2013-01-30

Corpus text ngram

Media type	textNgram	
Ngram	Base item	Other
	Order	1
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
Modality	Modality type	Written language
Size	1,202,209,147 words	

8.35. Bulgarian Automatic Collocations Dictionary

General Information

Short name	BACD
Description	The Bulgarian Automatic Collocations Dictionary produced by Lexical Computing Ltd., uses the web component of the Bulgarian National Corpus (419 million words, as prepared, lemmatised and tagged by the Institute for the Bulgarian Language (IBL), Sofia, Bulgaria). The sketch grammar was also as prepared by IBL. The dictionary has entries for 31,011 headwords, with an average of 4.2 collocations per headword. The entry for each

	collocation includes pointers to its corpus examples on the Sketch Engine website.
Identifier	835
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Other
Version	1.0
Last update	2013-01-30

Contacts

Adam Kilgarriff	
Position	Director
Contact	71, Freshfield Road adam@lexmasterclass.com http://www.sketchengine.co.uk/

Distribution

Availability	Available – unrestricted use	
IPR holder	Lexical Computing Ltd.	
	Contact	71, Freshfield Road BN2 0BL Brighton inquiries@sketchengine.co.uk http://www.sketchengine.co.uk

Licences

CC-BY-NC		
Restrictions of use	Attribution	
	Share alike	
Access medium	Downloadable	
Download location	http://gdex.sketchengine.co.uk/cesar/bg.zip	
Attribution text	Licence for Automatic Collocations DictionaryThis work is licensed under version 3.0 of the Creative Commons CC-BY-SA license as specified at http://creativecommons.org/licenses/by-sa/3.0/legalcode for the legal code of the license. This specifies that you are free to share (copy, distribute and transmit) the work and also to adapt it, provided that you acknowledge Lexical Computing Ltd. as creators (and give its website http://www.sketchengine.co.uk) (without suggesting that the company endorses your work). If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.	
Signatories	Lexical Computing Ltd.	
	Contact	71, Freshfield Road

		BN2 0BL Brighton inquiries@sketchengine.co.uk http://www.sketchengine.co.uk/
--	--	---

Metadata

Creation date	2013-01-31	
Metadata creators	Tsvetana Dimitrova	
	Position	Assistant Professor
	Contact	cvetana@dcl.bas.bg
	Organization	Institute for Bulgarian Language Department of Computational Linguistics ibl@ibl.bas.bg
Metadata last date updated	2013-02-01	

Resource creation

Creation end date	2012-10-14
--------------------------	------------

Lexical conceptual resource

Lexical conceptual resource type	Other
---	-------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
	Language script	Cyrl
Size	31,011 words	

8.36. Bibliography of Bulgarian Lexicology, Phraseology and Lexicography

General Information

Short name	BBLLPh
Description	The database of Bibliography of Bulgarian Lexicology, Phraseology and Lexicography contains bibliographic units for publications in Bulgarian lexicography, phraseology and

	lexicography that have been published since 1950. The publications are found in Bulgarian and foreign periodicals, linguistic volumes and series, electronic bibliographic databases and catalogues, etc. The database contains about 6,600 bibliographic records (over 200 monographs, 29 volumes, 32 textbooks, 1,968 articles and studies in different collections, 3,007 articles in periodicals, 186 dissertations, 202 reviews of scientific papers, 261 reviews of dictionaries, etc.). The database includes publications in Bulgarian, Russian, Ukrainian, Belarusian, Polish, Czech, Slovak, Serbian, Romanian, German, French, English, Italian, Spanish and Portuguese. Each bibliographic unit is accompanied by keywords (from a list of about 250 keywords) reflecting the content of the publication. URLs of electronic publications are given as well.
Identifier	836
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Other
URL	http://infolex.ibl.bas.bg/biblrecis/home.seam
Version	1.0
Last update	2013-01-10

Contacts

Diana Blagoeva	
Position	Associate Professor
Contact	d.blagoeva@ibl.bas.bg
Organization	Institute for Bulgarian Language Department of Bulgarian Lexicology and Lexicography ibl@ibl.bas.bg

Distribution

Availability	Available – restricted use	
IPR holder	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Bulgarian Lexicology and Lexicography
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia d.blagoeva@ibl.bas.bg http://ibl.bas.bg/en/index.htm
Availability start date	2013-01-21	

Licences

CC-BY-NC-ND

Restrictions of use	Academic – non-commercial use No redistribution
Access medium	Accessible through interface
Execution location	http://infolex.ibl.bas.bg/biblrecis/home.seam

Metadata

Creation date	2013-01-31	
Metadata creators	Tsvetana Dimitrova	
	Position	Assistant Professor
	Contact	cvetana@dcl.bas.bg
	Organization	Institute for Bulgarian Language Department of Computational Linguistics cvetana@dcl.bas.bg
Metadata last date updated	2013-01-31	

Validation

Validated	True
------------------	------

Usage

Foreseen use	Human use
Actual uses	Human use

Resource creation

Resource creator	Institute for Bulgarian Language	
	Short name	IBL
	Department name	Department of Bulgarian Lexicology and Lexicography
	Contact	52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia ibl@ibl.bas.bg http://ibl.bas.bg/en/index.htm
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://cesar.nytud.hu
	Funding type	EU funds
	Start date	2011-02-01

	End date	2013-01-30
	Complex Study of Bulgarian Lexis and Phraseology as a Part of the Cultural Heritage and National Identity	
	Funding type	National funds
Creation start date	2009-12-17	

Resource documentation

Reports	Blagoeva, Diana, Sia Kolkovska. „Инфолекс” – информационен лингвистичен портал за целите на българската лексикология, фразеология и лексикография. – Littera et Lingua Electronic Journal of Humanities, 2012, Fall, pp. 1-10. ISSN: 1312-6172 http://www.slav.uni-sofia.bg/liljournal
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Other
---	-------

Texts

Media type	text	
Linguality type	Multilingual	
Multilinguality type	Multilingual single text	
Languages	Bulgarian	
	Language ID	bg
	Language script	Cyrl
	Russian	
	Language ID	ru
	Language script	Cyrl
	Ukrainian	
	Language ID	uk
	Language script	Cyrl
	Belarusian	
	Language ID	be
	Language script	Cyrl
	Polish	
	Language ID	pl

	Language script	Latn
	French	
	Language ID	fr
	Language script	Latn
	Italian	
	Language ID	it
	Language script	Latn
	Spanish	
	Language ID	es
	Language script	Latn
	Czech	
	Language ID	cs
	Language script	Latn
	Slovak	
	Language ID	sk
	Language script	Latn
	Serbian	
	Language ID	sr
	Romanian	
	Language ID	ro
	Language script	Latn
	German	
	Language ID	de
	Language script	Latn
	Portuguese	
	Language ID	pt
	Language script	Latn
Modality	Modality type	Written language
Size	6,600 entries	

Text format	text
Character encoding	UTF-8

9. LSIL resources

9.1. Slovak National Corpus prim-5.0

General Information

Short name	prim-5.0
Description	The Slovak National Corpus is a representative corpus of contemporary Slovak language written texts published since 1955 (1953 being the time of most recent substantial Slovak language orthography reform). The corpus is automatically lemmatised and MSD tagged. The documents are annotated with their genre, style and other bibliographic information. There are specialised subcorpora containing fiction, informational texts, professional texts, original Slovak fiction, texts written from 1955 to 1989, and a balanced subcorpus. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
Identifier	901
Resource type	Corpus
URL	http://korpus.juls.savba.sk/prim(2d)5(2e)0.html
Version	5.0
Last update	2011-02-01

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk

	http://korpus.juls.savba.sk/
Availability start date	2011-02-01

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	https://bonito.korpus.sk

Metadata

Creation date	2011-11-21	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications Human use
Actual uses	NLP applications
	Human use

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	719000000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph

	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.2. Corpus of Spoken Slovak

General Information

Short name	hovor
Description	The database of the Corpus of Spoken Slovak contains audio records of spontaneous and semi-prepared speech from the entire Slovak territory and their text transcripts. Specific characteristics of spoken language are selectively captured in the transcripts, such as irregular structure of an utterance, pronunciation variants, means of speech modulation, and the presence of non-linguistic elements. The Corpus of Spoken Slovak provides material for research and description of the real form of contemporary standard spoken Slovak.
Identifier	902
Resource type	Corpus
URL	http://korpus.juls.savba.sk/shk.html
Version	4.0
Last update	2012-07-16

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Katarína Gajdošová	
Position	researcher
Contact	Panská 26 81364 Bratislava

	katarinag@korpus.juls.savba.sk
--	--

Distribution

Availability	Available – unrestricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2012-07-16	

Licences

CC-BY-SA	
Restrictions of use	Attribution Share alike
Access medium	Accessible through interface
Execution location	http://korpus.sk:8086/oral
GFDL	
Restrictions of use	Attribution Share alike
Access medium	Accessible through interface
Execution location	http://korpus.sk:8086/oral
AGPL	
Restrictions of use	Attribution Share alike
Access medium	Accessible through interface
Execution location	http://korpus.sk:8086/oral

Metadata

Creation date	2011-11-21	
Metadata creators	Radovan Garabík	
	Position	researcher

	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications Human use
Actual uses	NLP applications
	Human use

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	2600000 tokens	
Character encoding	UTF-8	
Annotation	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

Audio recordings

Media type	audio	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Audio size	282 hours	
Audio formats	Audio/speex	
	Sampling rate	44100
	Compression	True

	Compression name	Flac
	Compression loss	False
	Number of tracks	1
	Audio/vorbis	
	Sampling rate	44100
	Compression	True
	Compression name	Ogg vorbis
	Compression loss	True
	Number of tracks	1
	Audio/vorbis	
	Sampling rate	48000
	Compression	True
	Compression name	Ogg vorbis
	Compression loss	True
	Number of tracks	1
	Audio/flac	
	Compression	True
	Compression name	Flac
	Compression loss	False
	Number of tracks	2
Annotation	Segmentation	
	Segmentation level	Utterance
	Speech annotation – orthographic transcription	
	Segmentation level	Word
	Speech annotation – phonetic transcription	
	Segmentation	Word

	level	
	Speech annotation – sound events	
	Speech annotation – sound to text alignment	
	Segmentation level	Utterance
	Speech annotation – speaker identification	
	Speech annotation – speaker turns	

9.3. Slovak Morphology Database

General Information

Description	Slovak Morphological Database is a database of lemmas and their inflected wordforms with MSD tags.
Identifier	903
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Other

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – unrestricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/

Licences

--

AGPL	
Restrictions of use	Other
Access medium	Downloadable
CC-BY-SA	
Restrictions of use	Other
Access medium	Downloadable
GFDL	
Restrictions of use	Other
Access medium	Downloadable

Metadata

Creation date	2011-11-21	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Lexical conceptual resource

Lexical conceptual resource type	Other
---	-------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Modality	Modality type	Written language
Size	2470000 entries	

9.4. Slovak-Czech Parallel Corpus (all)

General Information

Description	Parallel Slovak-Czech corpus is a corpus of sentence aligned texts. Corpus consists of two parts: the subcorpus of fiction and the free subcorpus. The Slovak texts are morphologically annotated and disambiguated using the system applied in the Slovak National Corpus, Czech texts are annotated with the morče tagger. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
Identifier	904
Resource type	Corpus
URL	http://korpus.sk/skcs.html
Version	2.0
Last update	2012-09-08

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2012-09-08	

Licences

Proprietary	
Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	http://korpus.juls.savba.sk:8097/

Metadata

Creation date	2013-01-22	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Languages	Slovak	
	Language ID	sk
	Czech	
	Language ID	cs
Size	6433000 sentences	
Character encoding	UTF-8	
Annotation	Alignment	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.5. Slovak-English Parallel Corpus (all)

General Information

Description	The corpus consists of parallel Slovak and English texts, with automatic lemmatization, morphological analysis (for Slovak), POS tagging (for English). The corpus consists of two parts – the subcorpus of “fiction” and the free subcorpus. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
Identifier	905
Resource type	Corpus
URL	http://korpus.juls.savba.sk/sken.html
Version	2.0
Last update	2012-09-08

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2012-09-08	

Licences

Proprietary	
Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface

Execution location	http://korpus.juls.savba.sk:8098/
---------------------------	---

Metadata

Creation date	2013-01-22	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Languages	Slovak	
	Language ID	sk
	English	
	Language ID	en
Size	10000000 sentences	
Character encoding	UTF-8	
Annotation	Alignment	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	

	Segmentation level	Word
--	---------------------------	------

9.6. Slovak Treebank

General Information

Description	Slovak Language Treebank consists of 50000 manually syntactically annotated sentences, using the Prague Dependency Treebank methodology (analytical level). Most of the sentences have been annotated by two independent annotators.
Identifier	906
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Other

Contacts

Mária Šimková	
Position	researcher
Contact	Panská 26 81364 Bratislava marias@korpus.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2011-01-01	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Other

Metadata

Creation date	2012-07-16	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications Human use
Actual uses	Human use

Lexical conceptual resource

Lexical conceptual resource type	Other
---	-------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Modality	Modality type	Written language
Size	50000 sentences	

9.7. Balanced Slovak Corpus prim-5.0-vyv

General Information

Short name	VYV-5.0
Description	VYV is a balanced corpus with respect to text type. It contains 1/3 fiction, 1/3 informational text, 1/3 professional text (including popular science). The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
Identifier	907
Resource type	Corpus

URL	http://korpus.juls.savba.sk/prim(2d)5(2e)0.html
Version	5.0
Last update	2011-02-01

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚĽŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2011-02-01	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	https://bonito.korpus.sk

Metadata

Creation date	2012-07-16	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Metadata language ID	en
Metadata last date updated	2013-02-01

Usage

Foreseen use	NLP applications Human use
Actual uses	NLP applications
	Human use

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	247000000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.8. Manually Annotated Slovak Corpus

General Information

Short name	MAK

Description	MAK is a manually lemmatized and morphosyntactically annotated corpus. It is used as a basis for NLP tools training (primarily POS tagger and lemmatizer). This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed. The organization provides the ability to train your own tools, by providing access to the computer cluster (on request).
Identifier	908
Resource type	Corpus
URL	http://korpus.juls.savba.sk/stats.html
Version	3.0
Last update	2008-06-22

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚĽŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2008-06-22	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	https://bonito.korpus.sk

Metadata

Creation date	2012-07-16
----------------------	------------

Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications Human use
Actual uses	NLP applications
	Human use

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	1207000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.9. Language model prim-5.0-sane

General Information

Description	This is a language model from the Slovak National Corpus. This model is a 733 million token collection. Language model is in the iARPA format, using witten-bell smoothing. It was created by theIRSTLM Toolkit. It is lowercased. The model has been released with the contribution of the EuroMatrixPlus project.
Identifier	909
Resource type	Tool/service
Tool/service type	Tool
URL	http://korpus.sk/prim(2d)5(2e)0(2f)models.html

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – unrestricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2012-02-01	

Licences

Restrictions of use	Attribution Share alike
Access medium	Downloadable
Download location	http://korpus.sk/prim(2d)5(2e)0(2f)models.html

Metadata

Creation date	2012-07-16	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications
Actual uses	NLP applications

Tool/service

Tool/service type	Tool	
Tool/service subtype	Language Model	
Language dependent	True	
Output	Slovak	
	Media type	text
	Language ID	sk

9.10. Language model prim-5.0-inf

General Information

Description	This is a language model of journalistic style. The model is built on corpus of 515 million tokens. The language model is in iARPA format, using witten-bell smoothing. It was created by the IRSTLM Toolkit. The model is lowercased. It has been released with the contribution of the EuroMatrixPlus project.
Identifier	910
Resource type	Tool/service
Tool/service type	Tool
URL	http://korpus.sk/prim(2d)5(2e)0(2f)models.html

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – unrestricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚĽŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2012-02-01	

Licences

Restrictions of use	Attribution Share alike
Access medium	Downloadable
Download location	http://korpus.sk/prim(2d)5(2e)0(2f)models.html

Metadata

Creation date	2012-07-16	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications
Actual uses	NLP applications

Tool/service

Tool/service type	Tool	
Tool/service subtype	Language Model	
Language dependent	True	
Output	Slovak	
	Media type	text
	Language ID	sk

9.11. Language model prim-5.0-vyv

General Information

Description	This is a language model of balanced language. The model is built on the balanced Slovak corpus of 247 million tokens. The language model is in iARPA format, using witten-bell smoothing. It was created by the IRSTLM Toolkit. The model is lowercased. It has been released with the contribution of the EuroMatrixPlus project.
Identifier	911
Resource type	Tool/service
Tool/service type	Tool
URL	http://korpus.sk/prim(2d)5(2e)0(2f)models.html

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – unrestricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus

	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2012-02-01	

Licences

Restrictions of use	Attribution Share alike
Access medium	Downloadable
Download location	http://korpus.sk/prim(2d)5(2e)0(2f)models.html

Metadata

Creation date	2012-07-16	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications
Actual uses	NLP applications

Tool/service

Tool/service type	Tool	
Tool/service subtype	Language Model	
Language dependent	True	
Output	Slovak	
	Media type	text
	Language ID	sk

9.12. Corpus of Legal Texts

General Information

Short name	legal
Description	The corpus has been prepared in collaboration with the Ministry of Justice of the Slovak Republic. It is comprised of legal regulations and other available legal documents (laws, decrees, announcements, directives, protocols, etc.).
Identifier	912
Resource type	Corpus
URL	http://korpus.juls.savba.sk/extra.html
Version	1.0
Last update	2011-07-14

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚĽŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
	The Ministry of Justice of the Slovak Republic	
	Short name	MS SR
	Contact	Župné nám. 13 81311 Bratislava Milos.Matusek@justice.sk
Availability start date	2011-07-14	

Licences

Restrictions of use	Other
Access medium	Accessible through interface
Execution location	https://bonito.korpus.sk

Metadata

Creation date	2012-07-16	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications Human use
Actual uses	NLP applications
	Human use

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	146000000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence

	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word
Time coverages	1918-2011	

9.13. Slovak Web Corpus

General Information

Short name	sk-web
Description	Web corpus contains texts downloaded from the .sk domain. The texts are automatically lemmatized and morphologically tagged.
Identifier	913
Resource type	Corpus
URL	http://korpus.juls.savba.sk/extra.html
Version	2.0
Last update	2012-03-28

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚĽŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26

	81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2012-03-28

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	https://bonito.korpus.sk

Metadata

Creation date	2012-07-16	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications Human use
Actual uses	NLP applications
	Human use

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	1045000000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	

	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.14. Slovak-Czech Parallel Corpus (free)

General Information

Description	Parallel Slovak-Czech corpus is a corpus of sentence aligned texts of freely downloadable texts. The Slovak texts are morphologically annotated and disambiguated using the system applied in the Slovak National Corpus, Czech texts are annotated with the morče tagger.
Identifier	914
Resource type	Corpus
URL	http://korpus.sk/skcs.html
Version	2.0
Last update	2012-09-08

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ

	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2012-09-08	

Licences

Proprietary	
Restrictions of use	Academic – non-commercial use
Access medium	Downloadable
Execution location	http://korpus.juls.savba.sk/moses.html

Metadata

Creation date	2013-01-22	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Languages	Slovak	
	Language ID	sk
	Czech	
	Language ID	cs
Size	5700000 sentences	
Character encoding	UTF-8	

Annotation	Alignment	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.15. Slovak-English Parallel Corpus (free)

General Information

Description	The corpus consists of parallel Slovak and English freely downloadable texts, with automatic lemmatization, morphological analysis (for Slovak), POS tagging (for English). The corpus consists of original English language books and their Slovak translations.
Identifier	915
Resource type	Corpus
URL	http://korpus.juls.savba.sk/sken.html
Version	2.0
Last update	2012-09-08

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use
IPR holder	Jazykovedný ústav Ľudovíta Štúra

	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2012-09-08	

Licences

Proprietary	
Restrictions of use	Academic – non-commercial use
Access medium	Downloadable
Execution location	http://korpus.juls.savba.sk/moses.html

Metadata

Creation date	2013-01-22	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Languages	Slovak	
	Language ID	sk
	English	
	Language ID	en
Size	6000000 sentences	

Character encoding	UTF-8	
Annotation	Alignment	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.16. Slovak Terminology Database

General Information

Description	Slovak Terminology Database is a database of 6 000 terms from 23 fields.
Identifier	916
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Terminological resource
URL	https://data.juls.savba.sk/std/

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – unrestricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ

	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/

Licences

AGPL	
Restrictions of use	Other
Access medium	Downloadable
Execution location	http://data.juls.savba.sk/std/
CC-BY-SA	
Restrictions of use	Other
Access medium	Downloadable
Execution location	http://data.juls.savba.sk/std/
GFDL	
Restrictions of use	Other
Access medium	Downloadable
Execution location	http://data.juls.savba.sk/std/

Metadata

Creation date	2013-01-22	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Lexical conceptual resource

Lexical conceptual resource type	Terminological resource
---	-------------------------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Modality	Modality type	Written language
Size	6000 entries	

9.17. Corpus of Informational Texts prim-6.0-inf

General Information

Short name	prim-6.0-inf
Description	INF is a corpus consisting of informational text (mostly newspapers and journals). This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
Identifier	917
Resource type	Corpus
URL	http://korpus.juls.savba.sk/prim(2d)6(2e)0.html
Version	6.0
Last update	2013-01-22

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚĽŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/

Availability start date	2013-01-11
--------------------------------	------------

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	https://bonito.korpus.sk

Metadata

Creation date	2013-01-22	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications Human use
Actual uses	NLP applications
	Human use

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	889000000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	

	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.18. Corpus of Professional Texts prim-6.0-prf

General Information

Short name	prim-6.0-prf
Description	PRF is a corpus consisting of professional text (including popular science). The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
Identifier	918
Resource type	Corpus
URL	http://korpus.juls.savba.sk/
Version	6.0
Last update	2013-01-22

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLEŠ
	Department name	Slovenský národný korpus Slovak National Corpus

	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2013-01-11	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	https://bonito.korpus.sk

Metadata

Creation date	2013-01-22	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications Human use
Actual uses	NLP applications
	Human use

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	106000000 tokens	
Character encoding	UTF-8	

Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.19. Corpus of Fiction prim-6.0-img

General Information

Short name	IMG-6.0
Description	IMG is a corpus of fiction. The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
Identifier	919
Resource type	Corpus
URL	http://korpus.juls.savba.sk/prim(2d)6(2e)0.html
Version	6.0
Last update	2013-01-22

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use
---------------------	----------------------------

IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2013-01-11	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	https://bonito.korpus.sk

Metadata

Creation date	2013-01-22	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications Human use
Actual uses	NLP applications
	Human use

Texts

Media type	text
Linguality type	Monolingual

Languages	Slovak	
	Language ID	sk
Size	114000000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.20. Corpus of Original Slovak Texts prim-6.0-sk

General Information

Description	This is a corpus of original Slovak texts (no translations). The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
Identifier	920
Resource type	Corpus
URL	http://korpus.juls.savba.sk/prim(2d)6(2e)0.html
Version	6.0
Last update	2013-01-22

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava
	radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2013-01-11	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	https://bonito.korpus.sk

Metadata

Creation date	2013-01-22	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications
	Human use
Actual uses	NLP applications
	Human use

Texts

--	--

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	905000000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.21. Corpus of Original Slovak Fiction skimg-6.0

General Information

Short name	skimg-6.0
Description	Skimg-6.0 is a corpus of original Slovak fiction (no translations). The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
Identifier	921
Resource type	Corpus
URL	http://korpus.juls.savba.sk/prim(2d)6(2e)0.html
Version	6.0
Last update	2013-01-22

Contacts

Radovan Garabik	
Position	researcher

Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
----------------	---

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2013-01-11	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	https://bonito.korpus.sk

Metadata

Creation date	2013-01-22	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications Human use

Actual uses	NLP applications	
	Human use	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	35000000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmaization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.22. Corpus of Slovak Texts from the Years 1955 to 1989 R55AZ89

General Information

Short name	R55AZ89
Description	R55AZ89 is a corpus containing texts from the years 1955 to 1989. The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
Identifier	922
Resource type	Corpus
URL	http://korpus.juls.savba.sk/prim(2d)6(2e)0.html
Version	3.0
Last update	2013-01-22

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2013-01-11	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	https://bonito.korpus.sk

Metadata

Creation date	2013-01-22	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications Human use
Actual uses	NLP applications Human use

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	63000000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.23. Corpus of Historical Slovak

General Information

Short name	hist-1.0
Description	Corpus of Historical Slovak contains texts from the 16th, 17th and 18th centuries. The corpus is a database of electronically processed texts published in <i>Pramene k dejinám slovenčiny, I. - III;</i> (Sources for the History of Slovak).
Identifier	923
Resource type	Corpus

URL	http://korpus.juls.savba.sk/extra.html
Version	1.0
Last update	2012-12-30

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚĽŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2012-12-30	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	https://bonito.korpus.sk

Metadata

Creation date	2013-01-22	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Metadata language ID	en
Metadata last date updated	2013-01-30

Usage

Foreseen use	NLP applications Human use
Actual uses	NLP applications
	Human use

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	370000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word

9.24. Lietuvių kalbos WordNet (Lithuanian WordNet)

General Information

Description	Lietuvių kalbos WordNet projekto tikslas - aprašyti dažniausiai vartojamų lietuvių kalbos žodžių semantinius ryšius, remiantis anglų kalbos WordNet. Lithuanian WordNet is a lexical database including information about semantic relations of Lithuanian words. It is aligned with the Princeton 3.0 WordNet.
Identifier	924
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Wordnet

URL	http://korpus.sk/ltskwn_lt.html
------------	---

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – unrestricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/

Licences

AGPL	
Restrictions of use	Other
Access medium	Downloadable
Execution location	http://korpus.sk/ltskwn.html
CC-BY-SA	
Restrictions of use	Other
Access medium	Downloadable
Execution location	http://korpus.sk/ltskwn.html
Restrictions of use	Other
Access medium	Downloadable
Execution location	http://korpus.sk/ltskwn.html

Metadata

Creation date	2013-01-22
Metadata creators	Radovan Garabík

	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-02-05	

Lexical conceptual resource

Lexical conceptual resource type	Wordnet
---	---------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Lithuanian	
	Language ID	lt
Modality	Modality type	Written language
Size	12815 entries	

9.25. Slovník slovných spojení v slovenčine. Podstatné mená (Dictionary of Slovak Collocations. Nouns)

General Information

Description	Dictionary of noun collocations. It is the only one existing collocation dictionary in Slovakia. It is a dictionary of not only phrasemes, but also of common word collocations.
Identifier	925
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Other

Contacts

Peter Ďurčo	
Position	researcher
Contact	Námestie J. Herdu č. 2 917 01 Trnava durco@vronk.net

Distribution

Availability	Available – restricted use	
IPR holder	Filozofická fakulta Univerzity sv. Cyrila a Metoda v Trnave	
	Short name	FF UCM
	Department name	Katedra germanistiky
	Contact	Námestie J. Herdu č. 2 917 01 Trnava daniela.drinkova@ucm.sk http://kger.ff.ucm.sk/

Licences

Restrictions of use	Academic – non-commercial use
Execution location	http://vronk.net/wicol

Metadata

Creation date	2013-01-22	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Lexical conceptual resource

Lexical conceptual resource type	Other
----------------------------------	-------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk

Modality	Modality type	Written language
Size	250 entries	

9.26. Slovník slovných spojení v slovenčine. Prídavné mená (Dictionary of Slovak Collocations. Adjectives)

General Information

Description	Dictionary of adjective collocations.
Identifier	926
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Other

Contacts

Peter Ďurčo	
Position	researcher
Contact	Námestie J. Herdu č. 2 917 01 Trnava durco@vronk.net

Distribution

Availability	Available – restricted use	
IPR holder	Filozofická fakulta Univerzity sv. Cyrila a Metoda v Trnave	
	Short name	FF UCM
	Department name	Katedra germanistiky
	Contact	Námestie J. Herdu č. 2 917 01 Trnava daniela.drinkova@ucm.sk http://kger.ff.ucm.sk/

Licences

Restrictions of use	Academic – non-commercial use
Execution location	http://vronk.net/wicol

Metadata

Creation date	2013-01-22
----------------------	------------

Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Lexical conceptual resource

Lexical conceptual resource type	Other
---	-------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Modality	Modality type	Written language
Size	250 entries	

9.27. Slovak WordNet

General Information

Description	Slovak WordNet is a lexical database including information about semantic relations of Slovak words. It is aligned with the Princeton 3.0 WordNet.
Identifier	927
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Wordnet
URL	http://korpus.juls.savba.sk/WordNet.html

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava

	radovan.garabik@kassiopeia.juls.savba.sk
--	--

Distribution

Availability	Available – unrestricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/

Licences

AGPL	
Restrictions of use	Other
Access medium	Downloadable
Execution location	http://korpus.juls.savba.sk/WordNet.html
CC-BY-SA	
Restrictions of use	Other
Access medium	Downloadable
Execution location	http://korpus.juls.savba.sk/WordNet.html
Restrictions of use	Other
Access medium	Downloadable
Execution location	http://korpus.juls.savba.sk/WordNet.html

Metadata

Creation date	2013-01-22	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date	2013-01-30	

updated	
---------	--

Lexical conceptual resource

Lexical conceptual resource type	Wordnet
----------------------------------	---------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Modality	Modality type	Written language
Size	28800 entries	

9.28. Slovak National Corpus prim-6.0

General Information

Short name	prim-6.0
Description	The Slovak National Corpus is a representative corpus of contemporary Slovak language written texts published since 1955 (1953 being the time of most recent substantial Slovak language orthography reform). The corpus is automatically lemmatised and MSD tagged. The documents are annotated with their genre, style and other bibliographic information. There are specialised subcorpora containing fiction, informational texts, professional texts, original Slovak fiction, texts written from 1955 to 1989, and a balanced subcorpus. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
Identifier	928
Resource type	Corpus
URL	http://korpus.juls.savba.sk/prim(2d)6(2e)0.html
Version	6.0
Last update	2013-01-11

Contacts

Radovan Garabik	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚĽŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2013-01-11	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	https://bonito.korpus.sk

Metadata

Creation date	2011-11-21	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications Human use
Actual uses	NLP applications
	Human use

Texts

Media type	text

Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	1155000000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.29. Balanced Slovak Corpus prim-6.0-vyv

General Information

Short name	VYV-6.0
Description	VYV is a balanced corpus with respect to text type. It contains 1/3 fiction, 1/3 informational text, 1/3 professional text (including popular science). The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
Identifier	929
Resource type	Corpus
URL	http://korpus.juls.savba.sk/prim(2d)6(2e)0.html
Version	6.0
Last update	2013-01-11

Contacts

Radovan Garabík	
Position	researcher

Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
----------------	---

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚĽŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2013-01-11	

Licences

Restrictions of use	Academic – non-commercial use
Access medium	Accessible through interface
Execution location	https://bonito.korpus.sk

Metadata

Creation date	2012-07-16	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-30	

Usage

Foreseen use	NLP applications Human use
Actual uses	NLP applications

	Human use
--	------------------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	313000000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation - below POS tagging	
	Segmentation level	Word

9.30. Language model prim-6.0-sane

General Information

Description	This is a language model from the Slovak National Corpus. This lowercased model is from a 1200 million token collection. Language model is in the iARPA format, using witten-bell smoothing, created by theIRSTLM Toolkit.
Identifier	930
Resource type	Tool/service
Tool/service type	Tool
URL	http://korpus.sk/prim(2d)6(2e)0(2f)models.html

Contacts

Radovan Garabik

Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – unrestricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2013-01-24	

Licences

Restrictions of use	Attribution Share alike
Access medium	Downloadable
Download location	http://korpus.sk/prim(2d)6(2e)0(2f)models.html

Metadata

Creation date	2012-07-16	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-31	

Usage

--	--

Foreseen use	NLP applications
Actual uses	NLP applications

Tool/service

Tool/service type	Tool	
Tool/service subtype	Language Model	
Language dependent	True	
Output	Slovak	
	Media type	text
	Language ID	sk

9.31. Language model prim-6.0-inf

General Information

Description	This is a lowercased language model of journalistic style. The model is built on corpus of 889 million tokens. The language model is in iARPA format, using witten-bell smoothing and was created by the IRSTLM Toolkit.
Identifier	931
Resource type	Tool/service
Tool/service type	Tool
URL	http://korpus.sk/prim(2d)6(2e)0(2f)models.html

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – unrestricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚĽŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26

	81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2013-01-24

Licences

Restrictions of use	Attribution Share alike
Access medium	Downloadable
Download location	http://korpus.sk/prim(2d)6(2e)0(2f)models.html

Metadata

Creation date	2012-07-16
Metadata creators	Radovan Garabík
	Position researcher
	Contact Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en
Metadata last date updated	2013-01-31

Usage

Foreseen use	NLP applications
Actual uses	NLP applications

Tool/service

Tool/service type	Tool
Tool/service subtype	Language Model
Language dependent	True
Output	Slovak
	Media type text
	Language ID sk

9.32. Language model prim-6.0-vyv

General Information

Description	This is a lowercased language model of balanced language. The model is built on the balanced Slovak corpus of 313 million tokens. The language model is in iARPA format, using witten-bell smoothing, and was created by the IRSTLM Toolkit.
Identifier	932
Resource type	Tool/service
Tool/service type	Tool
URL	http://korpus.sk/prim(2d)6(2e)0(2f)models.html

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – unrestricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2013-01-24	

Licences

Restrictions of use	Attribution Share alike
Access medium	Downloadable
Download location	http://korpus.sk/prim(2d)6(2e)0(2f)models.html

Metadata

Creation date	2012-07-16	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2013-01-31	

Usage

Foreseen use	NLP applications
Actual uses	NLP applications

Tool/service

Tool/service type	Tool	
Tool/service subtype	Language Model	
Language dependent	True	
Output	Slovak	
	Media type	text
	Language ID	sk

9.33. Parallelum Slovaco-Latinum Corpus

General Information

Description	Parallel Slovak-Latin Corpus is a database of Latin texts translated into Slovak. Parallelum Slovaco-Latinum Corpus varios linguales textos utrius linguarum continet (id est Latinos textos interpretatos in lingua slovaca).
Identifier	933
Resource type	Corpus
URL	http://korpus.sk/skla.html
Version	2012-12-11
Last update	2012-12-11

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/

Metadata

Creation date	2013-01-30	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata last date updated	2013-01-30	

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Languages	Slovak	
	Language ID	sk
	Latin	
	Language ID	la

Modality	Modality type	Written language
Size	25000 sentences	
Character encoding	UTF-8	

9.34. n-grams from Slovak National Corpus

General Information

Description	Set of n-grams extracted from the Slovak National Corpus for $1 \leq n \leq 4$. The resource contains all unique n-grams preceeded and sorted by number of occurrences. There are separate files for case sensitive and for lowercased tokens.
Identifier	934
Resource type	Corpus
URL	http://korpus.juls.savba.sk/prim(2d)6(2e)0(2f)frequencies4.html http://korpus.juls.savba.sk/prim(2d)6(2e)0(2f)frequencies3.html http://korpus.juls.savba.sk/prim(2d)6(2e)0(2f)frequencies2.html http://korpus.juls.savba.sk/prim(2d)6(2e)0(2f)frequencies.html

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – unrestricted use
---------------------	------------------------------

Licences

CC-BY	
Access medium	Downloadable

Metadata

Creation date	2013-01-30	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava

	radovan.garabik@kassiopeia.juls.savba.sk
Metadata language name	English
Metadata language ID	en
Metadata last date updated	2013-01-30

Corpus text ngram

Media type	textNgram	
Ngram	Base item	Word
	Order	4
	Is factored	False
	Smoothing	Knesser-Ney
	Interpolated	False
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	686829767 4 – grams, 125027276 bigrams, 6268207 unigrams, 421417169 trigrams	

9.35. Multilingual Glossary of Synsets

General Information

Description	Multilingual glossary of synsets has been created by mapping several existing WordNets to the Princeton WordNet v. 3.0. It contains synsets (nouns and adjectives) in Bulgarian, Croatian, Hungarian, Serbian and Slovak, together with the links to the English WordNet.
Identifier	935
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Wordnet
URL	http://korpus.juls.savba.sk/SynsetGlossary_en.html
Version	2013-01-29

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26

	81364 Bratislava radovan.garabik@kassiopeja.juls.savba.sk
--	--

Distribution

Availability	Available – unrestricted use	
IPR holder	Jazykovedný ústav Ľudovíta Štúra	
	Short name	JÚLŠ
	Department name	Slovenský národný korpus Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
	Magyar Tudományos Akadémia, Nyelvtudományi Intézet	
	Department name	Nyelvtechnológiai és Élőnyelvi Osztály
	Contact	linginst@nytud.mta.hu http://www.nytud.hu
	Filozofski fakultet Sveučilišta u Zagrebu	
	Department name	Zavod za lingvistiku
	Contact	zzl@ffzg.hr http://www.ffzg.unizg.hr
	Универзитет у Београду, Математички факултет	
	Contact	matf@matf.bg.ac.rs http://www.matf.bg.ac.rs
	Институт за български език „Проф. Любомир Андрейчин”	
	Department name	Секция по компютърна лингвистика
	Contact	ibl@ibl.bas.bg http://www.ibl.bas.bg
	Szegedi Tudományegyetem, Informatikai Tanszékcsoport	
	Contact	depart@inf.u-szeged.hu http://www.inf.u-szeged.hu/
	MorphoLogic Kft.	
	Contact	proszeky@morphologic.hu http://www.morphologic.hu

Licences

--

Princeton_Wordnet	
Access medium	Downloadable

Metadata

Creation date	2013-01-30	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata last date updated	2013-01-31	

Lexical conceptual resource

Lexical conceptual resource type	Wordnet
---	---------

Texts

Media type	text	
Linguality type	Multilingual	
Multilinguality type	Other	
Languages	Slovak	
	Language ID	sk
	Bulgarian	
	Language ID	bg
	Croatian	
	Language ID	hr
	Serbian	
	Language ID	sr
	Language script	Cyrl
	Hungarian	
	Language ID	hu
Size	2296 entries	

9.36. Automatic Collocation Dictionary of Slovak

General Information

Description	The Automatic Collocations Dictionary of Slovak has been produced by Lexical Computing Ltd., based on the web corpus of Slovak language and Slovak word sketches. The entry for each collocation includes pointers to its corpus examples on the Sketch Engine website.
Identifier	936
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Machine readable dictionary
URL	http://korpus.juls.savba.sk/AutomaticCollocations_en.html

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – unrestricted use	
IPR holder	Lexical Computing Ltd.	
	Contact	71, Freshfield Road BN2 0BL Brighton inquiries@sketchengine.co.uk http://www.sketchengine.co.uk/
Availability end date	2012-10-26	

Licences

Access medium	Downloadable
----------------------	--------------

Metadata

Creation date	2013-01-31	
Metadata creators	Radovan Garabík	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Metadata language name	English
Metadata language ID	en
Metadata last date updated	2013-01-31

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary
---	-----------------------------

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Slovak	
	Language ID	sk
Size	47698 entries	