



CESAR

Central and Southeast European Resources

CIP-Pilot actions project no. 271022

www.cesar-project.net



Deliverable D3.2
Second batch of resources:
documentation of the delivery

Version No. 1.0
2012-07-31

Document Information

Deliverable number:	D3.2
Deliverable title:	Second batch of resources complying with the project's technical, linguistic, legal, etc. specifications
Due date of deliverable:	2012-07-31
Actual date of deliverable:	2012-07-31
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, Ida Raffaelli, Krešo Šojat, Vanja Štefanec, Marko Tadić (FFZG)</p> <p>Łukasz Degórski, Katarzyna Głowińska, Michał Lenart, Leszek Manicki, Małgorzata Marciniak, Marcin Miłkowski, Adam Przepiórkowski, Agata Savary, Filip Skwarski, Joanna Świetlicka, Zygmunt Vetulani, Adam Wardyński, Jakub Waszczuk, Dawid Weiss, Marcin Woliński, Bartosz Zaborowski (IPIPAN)</p> <p>Piotr Pęzik, Łukasz Drózdź, Maciej Buczek (ULodz)</p> <p>Cvetana Krstev, Ranka Stanković, Miloš Utvić, Duško Vitas (UBG)</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	Sta- tus	Author (Partner)	Contributions	Description/Approval Level
1.0	2012-07-31		Maciej Ogrodniczuk	All partners (listed above as participants)	Automatically generated content basing on metadata descriptions contributed by partners.

EXECUTIVE SUMMARY

This deliverable provides a documentation of the delivery of resources uploaded by CESAR consortium partners to META-SHARE CESAR node in the second batch delivery (July 2012). Due to possible updates the document also contains descriptions of resources from the first batch (November 2011).

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.....	8
1.1. Szeged Corpus.....	8
1.2. Szeged Treebank	9
1.3. Szeged Named Entity Recognition Corpus	11
1.4. Hungarian WordNet	12
1.5. Hungarian Webcorpus	13
1.6. Hunglish Corpus.....	14
1.7. morphdb.hu.....	16
1.8. hunmorph.....	17
1.9. hunalign	18
1.10. huntoken	20
1.11. Hungarian Opinion-Tagged Sentence Bank	21
1.12. HunNERwiki: Automatically generated NE tagged corpus for Hungarian.....	27
1.13. Hungarian Verb Phrase Constructions	30
1.14. hunner	32
1.15. hunpars	34
1.16. hunpos	36
1.17. Hungarian WSD Corpus.....	38
1.18. Hungarian Language Processing Tools in NooJ.....	41
1.19. SzegedParalell	43
1.20. Szeged Criminal NE Corpus	46
1.21. SzegedParalellFX	48
1.22. Szeged Treebank FX	50
2. BME-TMIT resources.....	52
2.1. Mindentudás Speech Corpus	52
2.2. Word level speech database for Hungarian	53
2.3. Hungarian BABEL	59
2.4. Hungarian Broadcast News Database.....	63
2.5. Sound Gesture Database.....	67
2.6. Hungarian Speech Emotion Database	70
2.7. Hungarian MTBA.....	74
2.8. Hungarian MRBA	78
2.9. Hungarian Phone Speech Call Center Database.....	82
2.10. Hungarian BABEL phonetic segmentation and syntactic and prosodic analysis	86
2.11. Di-phone database for text-to-speech conversion.....	89
2.12. Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection	94
2.13. Read speech database in Hungarian	100
2.14. Hungarian Book (Egri csillagok/Eclipse of the Crescent Moon by Géza Gárdonyi) Reading Speech and Aligned.....	106
2.15. Hungarian Poem (János vitéz/John the Valiant by Sándor Petőfi) Reading Speech and Aligned	108
2.16. Hungarian Parliamentary Speech and Aligned Text Selection Database	111

3. FFZG resources	114
3.1. Croatian National Corpus	114
3.2. Croatian Morphological Lexicon.....	117
3.3. Croatian-English Parallel Corpus	120
3.4. Croatian Lemmatisation Server	123
3.5. Croatian Valency Lexicon	126
3.6. Croatian Web Corpus	129
3.7. Slovene Web Corpus	132
3.8. Croatian-English Parallel Web Corpus.....	135
3.9. South-East European Parallel Corpus.....	137
3.10. Croatian Dependency Treebank	141
3.11. Web Content Extractor	143
3.12. Collocation and Term Extractor	146
4. IPIPAN resources	150
4.1. Polish Sejm Corpus	150
4.2. PoliMorf Inflectional Dictionary	156
4.3. Polish WordNet	160
4.4. Polish Named Entity Recognition Tool.....	164
4.5. 1 million subcorpus of National Corpus of Polish.....	169
4.6. Polish Named Entity Gazetteer.....	175
4.7. LUNA.PL Corpus.....	180
4.8. LUNA-WOZ.PL Corpus	183
4.9. Morphosyntactic tagset converter for positional tagsets	185
4.10. Spejd.....	188
4.11. N-grams from balanced National Corpus of Polish.....	192
4.12. Distributable subcorpus of National Corpus of Polish	195
4.13. Morfeusz PoliMorf	200
4.14. Morfologik Inflectional Dictionary	204
4.15. Grammatical Lexicon of Polish Phraseology	208
4.16. Grammatical Lexicon of Polish Economical Phraseology	212
4.17. Grammatical Lexicon of Warsaw Urban Proper Names	216
4.18. Multilingual lexicon of toponyms	220
4.19. Polish Valence Dictionary	224
4.20. Summarizer.....	226
4.21. morfologik-stemming	228
4.22. Corpus of the Polish language of the 1960s	231
4.23. Shallow Grammar for the National Corpus of Polish.....	235
4.24. PANTERA.....	239
4.25. PolNet.....	244
5. ULodz resources	248
5.1. PELCRA Polish-English parallel corpora (CC-BY)	248
5.2. PELCRA Polish-English parallel corpora (CC-BY-NC)	257
5.3. PELCRA Polish spoken corpus (CC-BY-NC)	263

5.4. ECL Dictionaries	268
5.5. PELCRA EN Lemmatizer	272
5.6. PELCRA Language Detector	277
5.7. PELCRA Polish-English parallel corpus of literary works (CC-BY)	281
5.8. PELCRA multilingual parallel corpora (CC-BY)	287
5.9. OSW Polish-English parallel corpus (CC-BY-NC)	308
5.10. PELCRA time-aligned spoken corpus of Polish (CC-BY-NC).....	314
5.11. PELCRA WebLign crawler.....	321
5.12. PELCRA Word Aligned Corpora.....	326
6. UBG resources	330
6.1. Serbian Wordnet.....	330
6.2. Corpus of Contemporary Serbian.....	334
6.3. Serbian Lemmatized and PoS Annotated Corpus.....	338
6.4. French-Serbian Aligned Corpus	342
6.5. Multilingual Edition of Verne's Novel "Around the World in 80 Days".....	346
6.6. Organizing digitized material	352
6.7. English-Serbian Aligned Corpus	355
6.8. Serbian NooJ module	360
6.9. Serbian Morphological Dictionary (Multext-East).....	365
6.10. Morphosyntactically tagged Serbian version of Jules Verne's novel "Around the world in 80 days"	368
6.11. Bibliša: Aligned Collection Search Tool.....	373
6.12. Corpus of Contemporary Serbian Newspapers and Magazines	376
7. IBL resources	380
7.1. Bulgarian National Corpus	380
7.2. Bulgarian National Corpus Collocation service	382
7.3. Bulgarian Part-of-Speech Corpus	384
7.4. Bulgarian Sense-annotated Corpus.....	387
7.5. Bulgarian-X language Parallel Corpus	389
7.6. Bulgarian WordNet	393
7.7. WordNet web service	395
7.8. Bulgarian Spell Checker for Windows	398
7.9. Bulgarian Spell Checker Web Service	400
7.10. Bulgarian-X Language Parallel Corpus Collocation service.....	403
7.11. Lists of Bulgarian Multiword Expressions.....	405
7.12. Bulgarian Frequency Dictionary	407
7.13. Hydra - tool for developing wordnets.....	410
7.14. Chooser – annotation tool.....	412
7.15. Bulgarian Sentence splitter and Tokenizer.....	414
7.16. Web based infrastructure for Bulgarian data processing.....	416
8. LSIL resources	419
8.1. Slovak National Corpus.....	419
8.2. Corpus of Spoken Slovak	421

8.3. Slovak Morphology Database	425
8.4. Slovak-Czech Parallel Corpus	426
8.5. Slovak-English Parallel Corpus.....	428
8.6. Slovak Treebank.....	430
8.7. Balanced Slovak Corpus.....	432
8.8. Manually Annotated Slovak Corpus.....	434
8.9. Language model prim-5.0-sane	436
8.10. Language model prim-5.0-inf.....	438
8.11. Language model prim-5.0-vyv	439
8.12. Corpus of Legal Texts	441
8.13. Slovak Web Corpus.....	443

Introduction

This deliverable provides a documentation of the delivery of resources made available in META-SHARE by CESAR consortium partners until the second batch (July 2012). 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/projectdirs/hlt
Version	2.0

Contacts

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

Distribution

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

Licences

CLARIN_ACA-NC

Restrictions of use	Academic - non-commercial use	
Access medium	CD-ROM	
Signatories	Zoltán Alexin	
	Contact	alexin@inf.u-szeged.hu
	Organization	University of Szeged Department of Informatics

Metadata

Creation date	2011-10-17	
Metadata creators	Zoltán Alexin	
	Contact	alexin@inf.u-szeged.hu
	Organization	University of Szeged Department of Informatics

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

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

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

Distribution

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

Licences

CLARIN_ACA-NC	
Restrictions of use	Academic - non-commercial use
Access medium	CD-ROM
Signatories	Zoltán Alexin
	Contact alexin@inf.u-szeged.hu
	Organization University of Szeged Department of Informatics

Metadata

Creation date	2011-10-17
Metadata creators	Zoltán Alexin
	Contact alexin@inf.u-szeged.hu
	Organization University of Szeged Department of Informatics

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

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

Distribution

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

Licences

CLARIN ACA-NC		
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

Metadata

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

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/nlp?lang=en&page=nlpproj_hunont
Version	1.0

Contacts

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

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

Distribution

Availability	Under negotiation
---------------------	-------------------

Licences

UnderNegotiation	
Restrictions of use	Other
Access medium	CD-ROM

Metadata

Creation date	2011-10-17	
Metadata creators	Zoltán Alexin	
	Contact	alexin@inf.u-szeged.hu
	Organization	University of Szeged Department of Informatics

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 gigacorpous 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

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

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

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

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 level	Sentence

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

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

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

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

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

Tamás Prajczér	
Position	CEO
Contact	Bécsi út 126-128. 1034 Budapest prajczér@geox.hu http://geox.hu/ceginformacio/kapcsolat/prajczér
Organization	GeoX Térinformatikai Kft.
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

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 Commercial use Attribution	
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.
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
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
	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		
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.
		Kornél Koósz	
		Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/

	Organization	GeoX Térinformatikai Kft.
	Júlia Domán	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft.
	Andrea Koronkai	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft.
	Zsófia Csikár	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft.
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.
	Kornél Koósz	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft.
	Júlia Domán	
	Contact	Bécsi út 126-128. 1034 Budapest

		prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft.
	Andrea Koronkai	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft.
	Zsófia Csikár	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft.
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.
	Kornél Koósz	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft.
	Júlia Domán	
	Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
	Organization	GeoX Térinformatikai Kft.
	Andrea Koronkai	

		Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
		Organization	GeoX Térinformatikai Kft.
		Zsófia Csikár	
		Contact	Bécsi út 126-128. 1034 Budapest prajczer@geox.hu http://geox.hu/
		Organization	GeoX Térinformatikai Kft.

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

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

Distribution

Availability	Available – unrestricted use	
IPR holder	GeoX Térinformatikai Kft.	
	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	Share alike Attribution	
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
Distribution rights holder	GeoX Térinformatikai Kft.	
	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	Tibor Pinter

creators	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
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.

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	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, hundisambig	
	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
		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	

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 sg into consideration' has a lexically free direct object and a lexically fixed into-PP).
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

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
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

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
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-02	

Usage

Actual uses	NLP applications	
	Reports	Gábor Recski, 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

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
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-02	

Usage

Actual uses	NLP applications	
	Reports	Babarczy Aanna – Gábor B. – Hamp Gábor – Kárpáti A. – Rung András – Szakadát István 2005. Mondattani elemző alkalmazás, In: Alexin Zoltán – Csendes Dóra (szerk.), III.

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

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
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_hunwsd

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
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
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

Distribution

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

Licences

MSCommons_NoCOM-NC-NR		
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
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

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	14
Modality	Modality type	Other
Size	14	
Character encoding	UTF-8	
Annotation	Semantic annotation	
	Segmentation	Word

	level	
	Annotation mode	Automatic

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 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</p>
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

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
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

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
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
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

Distribution

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

Licences

MSCommons_NoCOM-NC-NR		
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
Distribution	University of Szeged	

rights holder	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

Usage

Actual uses	NLP applications	
	Reports	Toth, Krisztina; Farkas, Richard; Kocsor, Andras 2006: Hybrid algorithm for sentence alignment of Hungarian-English parallel corpora. In: Conference of PhD students on Computer Sciences. Volume of Extended Abstracts, 27-30th June, Szeged, Hungary, pp. 97-98.

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	90 sentences

	English
	Language ID EN
	Size 90 sentences
Size	99 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
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

Distribution

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

Licences

MSCommons NoCOM-NC-NR	
Restrictions of use	Academic - non-commercial use
Access medium	Downloadable
Download	http://www.inf.u-szeged.hu/rgai/corpus_ne

location	
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
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

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	540 tokens
Modality	Modality type	Other
Size	540 tokens	
Annotation	Semantic annotation – named entities	

1.21. SzegedParalellFX

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

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
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
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

Distribution

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

Licences

MSCommons NoCOM-NC-NR	
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
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

Usage

Actual uses	NLP applications
	Reports Vincze, Veronika 2012: Light Verb Constructions in the SzegedParalellFX 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 CESAR

	name	
	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	14 sentences	
Annotation	Alignment	

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

Distribution

--	--

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

Licences

MSCommons_NoCOM-NC-NR		
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
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

Usage

Actual uses	NLP applications
	Reports

Resource creation

Funding projects	Central and South-East European Resources
	Project short

name	
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	82 sentences
	English	
	Language ID	EN
	Size	82 sentences
Size	82 sentences	
Annotation	Other	

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

Distribution

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

Licences

UnderNegotiation	
Restrictions of use	Other
Access medium	Downloadable
Download location	http://mindentudas.hu/media/mp3
Signatories	Dániel Varga
	Contact daniel@mokk.bme.hu
	Organization Budapest University of Technology and Economics Department of Telecommunications and Media Informatics
Distribution rights holder	Dániel Varga
	Contact daniel@mokk.bme.hu
	Organization Budapest University of Technology and Economics MOKK Centre for Media Research and Education

Metadata

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

Texts

Media type	text
Linguality type	Monolingual
Languages	Hungarian
	Language ID HU
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

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
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
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

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
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 Olaszy	
	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

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 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
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	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	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		
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		No		

		of persons	
		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

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
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
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
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	Natural numbers Isolated words 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 tracks	1	
	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 persons	no dialect
		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

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
Availability start date	2011-11-30	

Licences

MSCommons NoCOM-NC-NR-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
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

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
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 understanding Discourse analysis Speaker identification Speech to speech translation	
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
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 resource type	Lexicon
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

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
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
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 vicsi@tmit.bme.hu http://alpha.tmit.bme.hu/speech/
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics
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	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

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
Availability start date	2011-11-30	

Licences

MSCommons_NoCOM-NC-NR-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
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

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

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
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	20

		start	
		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	<p>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 dialectical, 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 dialectical 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.</p>
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

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
Availability start date	2012-07-02	

Licences

MSCommons_COM-NR-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
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2012-07-02	
Revision		

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Speech analysis Speech recognition Person 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
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	Natural numbers Isolated words 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 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 + 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</p>

	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 ortographical transcription of the spoken material. Almost one third of the database (100 speakers' utterances) was manually segmentated and labelled at phoneme level, using SAMPA codes.
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

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
Availability start date	2012-07-02	

Licences

MSCommons COM-NR-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
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2012-07-02	
Revision		

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Speech analysis Speech recognition Person 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
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	annotation based on listening

	mode details		
	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

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
Availability start date	2012-07-02	

Licences

MSCommons COM-NR-ND-FF	
Restrictions of use	No redistribution No derivatives
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 ortographic transcription.
MSCommons NoCOM-NC-NR-ND-FF	
Restrictions of use	No redistribution No derivatives
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 ortographic 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
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2012-07-02	
Revision		

Usage

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

	mode details		
	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

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
Availability start date	2012-08-01	

Licences

MSCommons NoCOM-NC-NR	
Restrictions of use	No redistribution
Access medium	Downloadable
Download location	http://alpha.tmit.bme.hu/speech/hdbbbabel.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

Source	CESAR
Metadata language ID	en-us
Metadata last date updated	2012-07-02
Revision	

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
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

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
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
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

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
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

Usage

Foreseen use	Human use NLP applications	
NLP-specific use	Speech synthesis Talking head synthesis Speech analysis	
Actual uses	NLP applications	
	NLP-specific use	Speech synthesis 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 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
Creation start date	2002-01-01	
Creation end date	2012-06-30	

Resource documentation

Reports	Olaszy G. - Gépi beszéd-keltés 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		
	Start date	2007-01-01		
	End date	2012-06-30		
	Size	10693 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	
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	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

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
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
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

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

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

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
	Tamás Bóhm	
	Position	Ph.D. lecturer
	Contact	Magyar tudósok körútja 2. H-1117 Budapest bohmf@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics
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	Naturalness	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 laklara@tmit.bme.hu http://www.tmit.bme.hu	
Organization		Budapest University of Technology and Economics	

			Department of Telecommunications and Media Informatics	
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		
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	corpora		

	source
--	--------

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

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
Availability start date	2012-07-15	

Licences

CLARIN RES	
Restrictions of	Academic - non-commercial use

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
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

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
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
	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
	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

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
	Tamás Bóhm	
	Position	Ph.D. lecturer
	Contact	Magyar tudósok körútja 2. H-1117 Budapest bohmf@tmit.bme.hu http://speechlab.tmit.bme.hu
	Organization	Budapest University of Technology and Economics Department of Telecommunications and Media Informatics
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 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	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	Klára Laczkó	
		Position	staff member
		Contact	Magyar tudósok körútja 2. H-1117 Budapest laklara@tmit.bme.hu http://www.tmit.bme.hu
Organization		Budapest University of Technology and Economics	

		Department of Telecommunications and Media Informatics		
		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		
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		
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 impairment of persons	No	
Speaking	No			

		impairment of persons	
		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 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	1.0

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

Distribution

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

Licences

CC BY	
Access medium	Downloadable
Download location	ftp://ask.contact.person

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

Metadata

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

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	Wave/audio	
	Signal encoding	Linear PCM
	Sampling rate	16000
	Quantization	16
	Compression	False
	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%.
	Annotation tool	voXerver 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 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	1.0

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

Distribution

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

Licences

CC BY		
Access medium	Downloadable	
Download location	ftp://ask.contact.person	
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

Metadata

Creation date	2012-07-12	
Metadata creators	András Balog	
	Position	engineer
	Contact	abalog@aitia.ai
	Organization	THINKTech Research Center non-profit LLC
Source	CESAR	
Metadata language ID	en-us	
Metadata last	2012-07-12	

date updated	
---------------------	--

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	Naturalness	Read speech
	Conversational type	Monologue
	Scenario type	Other
	Audience	No
	Interactivity	Non interactive
Audio formats	Wave/audio	
	Signal encoding	Linear PCM
	Sampling rate	16000
	Quantization	16
	Compression	False
	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%.
	Annotation tool	voXerver 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	1.0

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

Distribution

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

Licences

CC_BY		
Access medium	Downloadable	
Download location	ftp://ask.contact.person	
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

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
Source	CESAR	
Metadata language ID	en-us	
Metadata last date updated	2012-07-10	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Hungarian	
	Language ID	HU
	Size	9999 words
Size	9999 phonemes	
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
	Size	9999 seconds
Audio size	9999 bytes (9999 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	Wave/audio	
	Signal encoding	Linear PCM
	Sampling rate	16000
	Quantization	16
	Compression	False
	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 database comprises only those segments where the match between automatic and manual transcriptions

		is over 9999 %.
	Annotation tool	voXerver ASR engine, other self developed processing tools
Creation	Original source	http://www.parlament.hu/internet/plsql/ogy_naplo.naplo_ujnapok_ckl

3. FFZG resources

3.1. Croatian National Corpus

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://hnk.ffzg.hr/
Version	2.5.1
Last update	2011-09-01

Contacts

Marko Tadić	
Position	Head of the Chair for 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://hnk.ffzg.hr/

Licences

Proprietary		
Restrictions of use	Academic - non-commercial use	
Access medium	Accessible through interface	
Execution location	http://hnk2.ffzg.hr	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences, Department/Institute of Linguistics	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hnk.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, 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
	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; Osolobě, 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	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 – below POS tagging	
	Segmentation level	Word

3.2. Croatian Morphological Lexicon

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 Lemmatization 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 for 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	Marko Tadić	
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hnk.ffzg.hr
	University of Zagreb, Faculty of Humanities and Social Sciences, Department/Institute of Linguistics	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hnk.ffzg.hr

Licences

Proprietary		
Restrictions of use	Academic - non-commercial use	
Access medium	Accessible through interface	
Execution location	http://hml.ffzg.hr	
Fee	negotiable	
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences, Department/Institute of Linguistics	
	Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hnk.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 http://hnk.ffzg.hr/mt
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department/Institute of Linguistics
Metadata language ID	en	

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences,
-------------------------	---

	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 Own 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	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, Lisabon-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 Part of speech Case Gender

		Number Degree Mood Tense Person Auxiliary Inflection Other
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	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://hnk.ffzg.hr/hr-en_p-corp
Version	2
Last update	2011-09-01

Contacts

Marko Tadić	
Position	Head of the Chair for 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 – 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 zsl@ffzg.hr http://hnk.ffzg.hr/

Licences

CC_BY-NC-SA_3.0	
Restrictions of use	Academic - non-commercial use
Access medium	Downloadable
Download location	http://hnk.ffzg.hr/hr-en_p-corp/download/CW_v02_tmx.zip
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences, Department/Institute of Linguistics
	Contact Ivana Lučića 3 10000 Zagreb zsl@ffzg.hr http://hnk.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, 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	latin
	English

	Language ID	en
	Language script	latin
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	Lexical conceptual resource
Lexical conceptual resource type	Computational lexicon
URL	http://hml.ffzg.hr/
Version	2.0
Last update	2011-11-20

Contacts

--

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

Licences

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

Metadata

Creation date	2011-11-20	
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://hnk.ffzg.hr/mt
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department/Institute of Linguistics
Metadata language ID	en	

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences, 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 Own 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	2003-04-01	

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
----------------	---

Lexical conceptual resource

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

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	latin
Modality	Modality type	Written language
Size	4000000 words	
Character encoding	UTF-8	

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 Humanistics 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
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-odsjecka/190-nives-mikeli-preradovi-
Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department

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-odsjeka/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 3.0		
Restrictions of use	Academic - non-commercial use	
Access medium	Downloadable	
Download location	http://cal.ffzg.hr/crovallex/index.html	
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://hnk.ffzg.hr/mt
	Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department/Institute of Linguistics
Metadata language ID	en	

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-odsjecka/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 Own 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	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 ; Zaharim, 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	Encoding level	Syntax Semantics

resource encoding	Linguistic information	Lemma Part of speech Semantics – semantic roles Other
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	latin
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	311
Resource type	Corpus
URL	http://www.nljubesic.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.nljubestic.net/
Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department 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, Department of Information Sciences
	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
Execution location	http://www.nljubestic.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://hnk.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

Metadata language ID	en
Metadata last date updated	2012-07-30

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	2011-06-01	

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	latin
Size	1 200 000 000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation	Paragraph

	level	
	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	312
Resource type	Corpus
URL	http://www.nljubestic.net/resources/corpora/slwaC/
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

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://hnk.ffzg.hr/

Licences

CC BY-SA	
Restrictions of use	Attribution Share alike
Access medium	Downloadable
Execution location	http://www.nljubesic.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://hnk.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
Metadata language ID	en	
Metadata last date updated	2012-07-30	

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	2011-06-01	

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	latin
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 parallel 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 threshold 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	313
Resource type	Corpus
URL	http://www.nljubesic.net/resources/corpora/hrenwac/
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

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://hnk.ffzg.hr/

Licences

CC BY-SA

Restrictions of use	Attribution Share alike	
Access medium	Downloadable	
Execution location	http://www.nljubesic.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://hnk.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
Metadata language ID	en	
Metadata last date updated	2012-07-30	

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	2012-04-01
----------------------------	------------

Resource documentation

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

Texts

Media type	text	
Linguality type	Bilingual	
Languages	Croatian	
	Language ID	hr
	Language script	latin
	English	
	Language ID	en
	Language script	latin
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	314
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

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://hnk.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://hnk.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
Metadata language ID	en	
Metadata last date updated	2012-07-30	

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	2012-04-01	

Resource documentation

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

Texts

Media type	text	
Linguality type	Multilingual	
Languages	Albanian	
	Language ID	sq
	Language script	latin
	Bosnian	

	Language ID	bs
	Language script	latin
	Bulgarian	
	Language ID	bg
	Language script	cyrillic
	Croatian	
	Language ID	hr
	Language script	latin
	English	
	Language ID	en
	Language script	latin
	Greek	
	Language ID	el
	Language script	alphabet
	Macedonian	
	Language ID	mk
	Language script	cyrillic
	Romanian	
	Language ID	ro
	Language script	latin
	Serbian	
	Language ID	sr
	Language script	cyrillic
	Turkish	
	Language ID	tr
	Language script	latin
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	315
Resource type	Corpus
URL	http://hobs.ffzg.hr/
Version	1.0
Last update	2012-07-30

Contacts

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

Licences

CC BY-NC-SA	
Restrictions of use	Attribution Academic - non-commercial use Share alike

Access medium	Downloadable
Execution location	http://hobs.ffzg.hr
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences, 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	2012-07-30

Resource creation

Resource creator	University of Zagreb, Faculty of Humanities and Social Sciences, 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	2007-06-01

date	
-------------	--

Resource documentation

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

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Croatian	
	Language ID	hr
	Language script	latin
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 – below 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	316
Resource type	Tool/service
Tool/service type	Tool
URL	http://www.nljubesic.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.nljubesic.net/
Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences

Distribution

Availability	Available – unrestricted use	
IPR holder	Marko Tadić	
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hnk.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 zsl@ffzg.hr http://hnk.ffzg.hr/

Licences

ApacheLicence_V2.0	
Restrictions of use	Inform licensor

Access medium	Downloadable
Execution location	http://www.nljubesic.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://hnk.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
Metadata language ID	en
Metadata last date updated	2012-07-30

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

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	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	
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	316
Resource type	Tool/service
Tool/service type	Tool
URL	http://www.nljubesic.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.nljubesic.net/
Organization	University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences

Distribution

Availability	Available – unrestricted use	
IPR holder	Marko Tadić	
	Contact	Ivana Lučića 3 10000 Zagreb marko.tadic@ffzg.hr http://hnk.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://hnk.ffzg.hr/

Licences

ApacheLicence V2.0	
Restrictions of use	Inform licensor

Access medium	Downloadable
Execution location	http://www.nljubesic.net/resources/tools/collterm/
Distribution rights holder	University of Zagreb, Faculty of Humanities and Social Sciences
Contact	Ivana Lučića 3 10000 Zagreb zzl@ffzg.hr http://hnk.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
Metadata language ID	en
Metadata last date updated	2012-07-30

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	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
	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	

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	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	
Tool/service creation	Implementation language	Python	

4. IPIPAN 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

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
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 Own funds National 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	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	Automatic

mode	
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 practices	TEI
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 converted 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	<p>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 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 maciej.ogrodniczuk@ipipan.waw.pl http://zil.ipipan.waw.pl/MaciejOgrodniczuk</p>

Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group
---------------------	---

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
Source	CESAR	
Metadata	en	

language ID	
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 Own funds National 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	Zygmunt Saloni, Włodzimierz Gruszczyński, Marcin Woliński and Robert Wołosz. Słownik gramatyczny języka polskiego. Wiedza Powszechna. Warszawa 2007. Marcin Miłkowski. 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	Encoding level	Morphology

resource encoding	Linguistic information	Lemma Part of speech Case Gender Number Degree Mood Tense Person Aspect Auxiliary Inflection
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	

4.3. 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 103000 lexical units

	in 74000 synsets. The current version of the resource introduces 5500 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

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	Semantic role labelling Document classification

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	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
	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
	SyNaT — Research Task: “Construction of an open, repository hosting and communication platform for the network knowledge resources for 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 – relations Semantics – cross references 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	105074 synsets, 96276 lemmata (strings in Princeton WordNet), 145439 lexical units (i.e., lemma-sense pairs), 9283 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 Recognition (NER) based on the Conditional Random Fields (CRF) modelling method. The tool has been constructed as a part of the National Corpus of Polish project. It has been adapted to recognize tree-like structures of NEs (i.e., with recursively embedded NEs) using the Joined Label Tagging (JLT) method. The JLT method is a simple method of encoding NE structures as a sequence of labels. With this method various additional informations about NEs of categorical nature – type, subtype, type of derivation – can be encoded on the level of labels and subsequently recognized using the resultant CRF model. The tool can be configured to use various types of observations during the training and recognition process, for example: lexical informations from textual level, or grammatical informations from morphosyntactic level.
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

Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group
---------------------	---

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/Nerf?action=AttachFile&do=get&target=nerf.dist.0.2.tgz	
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	2011-11-24	
Metadata creators	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
	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
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
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group
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
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	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	Linux	
Required software	Python (version 2.6 or higher)	
	Cython (version 0.14 or higher)	

Tool/service evaluation	Evaluated	True
	Level	Diagnostic
	Type	Black box
	Criteria	Intrinsic
	Measure	Automatic
	Reports	Results (which will be described in the NKJP book; earlier evaluation results can be found in the "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) can be acquired by running <code>cross_validate.py</code> script distributed with the Nerf tool. The process is described in details in the README file, also distributed with Nerf.
	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
	Contact	Jana Kazimierza 5 01-248 Warsaw waszczuk.kuba@gmail.com
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group
Tool/service creation	Implementation language	Python Cython C
	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

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=get&target=NKJP-PodkorpusMilionowy-1.0.tgz
Fee	free of charge
Signatories	Adam Przepiórkowski
	Contact

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	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	TEI

to standards best practices	
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	TEI

	best practices	
	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	various shell scripts Anotatornia Morfeusz SGJP Nerf Spejd

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
---------------------	--

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	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

	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
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	Named entity recognition		

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

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. PISKORSKI, J. (2005). Named-Entity Recognition for Polish with SProUT, in LNCS Vol 3490: Proceedings of IMTCI 2004, Warsaw, Poland. 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.
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 Part of speech Case Gender Number Other
	Conformance to standards best practices	LMF
Creation	Original source	Publications of the Commission for Standardization of Geographic Names of Polish Frontiers (Komisja Standaryzacji Nazw Geograficznych poza Granic Rzeczypospolitej Polskiej), freely available at http://www.gugik.gov.pl/kom 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 etc.), historical regions of Europe (http://pl.wikipedia.org/wiki/Kategoria:Regiony_i_krainy_historyczne_Euro), 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 National Corpus of Polish".
Creation tools	SProUT Extraction Platform adapted to Polish Named Entity Annotation with fully automated rule-based NER system for Polish

Texts

Media type	text
Linguality type	Monolingual
Languages	Polish
	Language ID PL
	Language script latin
Size	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
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group

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
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
	Malgorzata Marciniak
	Contact Jana Kazimierza 5

	01-248 Warsaw malgorzata.marciniak@ipipan.waw.pl http://zil.ipipan.waw.pl/MalgorzataMarciniak
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group
Michał Lenart	
Position	Assistant Engineer
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
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	500 files, 12778 utterances, 81049 words	
Annotation	Segmentation	
	Segmentation level	Utterance Word Word group
	Semantic annotation	
	Segmentation level	Word group Clause

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
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group

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

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
	Malgorzata Marciniak	
	Position	Assistant Professor
	Contact	Jana Kazimierza 5 01-248 Warsaw malgorzata.marciniak@ipipan.waw.pl http://zil.ipipan.waw.pl/MalgorzataMarciniak
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group
	Michal Lenart	
	Position	Assistant Engineer
	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
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	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 IPIAN 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

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

type	
------	--

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	
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

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	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
	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
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	Shallow parsing of the National Corpus of Polish.		

details	
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 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

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
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group

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
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
	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
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
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group
Funding projects	Central and South-East European Resources	
	Project short name	CESAR
	URL	http://www.cesar-project.net
	Funding type	EU funds Own funds National 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	5364398 unigrams, 75395184 bigrams, 170180746 trigrams, 217586930 4 –	

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

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/DistrNKJP
Fee	free of charge
Signatories	Adam Przepiorkowski
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	2012-07-06	

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
Size	99280766 words	
Annotation	Segmentation	
	Annotation	True

standoff	
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 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
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	Nerf Spejd Pantera

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

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

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

Metadata

Creation date	2012-07-25
Metadata creators	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
	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
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
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. Wierzchoń 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://sgip.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	Segmentation	
	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

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
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 Own funds National 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.</p> <p>http://doi.wiley.com/10.1002/spe.971</p>
----------------	---

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology
	Linguistic information	<p>Lemma</p> <p>Part of speech</p> <p>Case</p> <p>Gender</p> <p>Number</p> <p>Degree</p> <p>Mood</p> <p>Tense</p> <p>Person</p> <p>Aspect</p> <p>Auxiliary</p> <p>Inflection</p>
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 wyrażeń 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 Nekst 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

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	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
	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
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	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
	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
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	<p>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.</p> <p>CZEREPOWICKA, M., KOSEK, I. (2011): Problemy opisu związków frazeologicznych w formalizmie „Multifleks” (na przykładzie rodzaju wyrażen frazeologicznych), in "Różne formy, różne treści", pp. 117–126, Warszawa 2011.</p> <p>CZEREPOWICKA, M. (2011): „Toposław” jako narzędzie znakowania jednostek wielocłonowych' in Matusiak-Kempa, I., Przybyszewski, S. (eds.) Nowe zjawiska w języku, tekście, komunikacji. Kontekst a komunikacja, Olsztyn, pp. 28–35.</p>
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology Syntax
	Linguistic information	Lemma Other Part of speech Case Gender Number Degree Mood Tense Person Aspect Auxiliary Inflection
Creation	Original source	<p>KOSEK, I. (2008): "Fleksja i składnia nieciągłych imiennych jednostek leksykalnych", Olsztyn.</p> <p>BAŃKO, M. (2004): "Słownik porównań", Wydawnictwo Naukowe PWN, Warszawa.</p> <p>CZEREPOWICKA, M. (2006): "Opis powierzchniowoskładniowy wyrażen niestandardowych typu na lewo, do dziś, po trochu, na zawsze we współczesnym języku polskim", Akademicka Oficyna Wydawnicza EXIT, Warszawa.</p> <p>WOJDAK, P. (2004): "Przysłówki polisegmentalne w modelu składniowym polszczyzny", Wydawnictwo Naukowe US, Szczecin.</p> <p>plWordNet</p>

		(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	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 wyrażeń Frazologicznych 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

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
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	Filip Makowiecki	
	Contact	f.makowiecki@gmail.com
	Organization	University of Warsaw
	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
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 Other Part of speech Case Gender Number Degree Mood Tense Person Aspect Auxiliary Inflection
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	11212 entries, 146861 multi-word units, 305 rules	
Character	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 WArzawy) 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

Małgorzata 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

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

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
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	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
	Celina Heliasz	
	Contact	celinaheliasz@op.pl
	Organization	University of Warsaw
	Joanna Rabięga-Wińniewska	
	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
	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	
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	2007-01-01	

date	
------	--

Resource documentation

Reports	<p>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.</p> <p>MARCINIAK, M., RABIEGA-WIŚNIEWSKA, 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.</p>
----------------	--

Lexical conceptual resource

Lexical conceptual resource type	Machine readable dictionary	
Lexical conceptual resource encoding	Encoding level	Morphology Syntax
	Linguistic information	Lemma Other Part of speech Case Gender Number Degree Mood Tense Person Aspect Auxiliary Inflection
Creation	Original source	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. Marciniak, M., Rabiega-Wiśniewska, 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	Mixed

	mode	
	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	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 and Romanian. 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	0.1
Last update	2012-07-24

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

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://bach.ipipan.waw.pl/redmine/attachments/116/wikipedia_translations.prolexbase	
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	Leszek Manicki	
	Contact	Wilczak 13/54 61-623 Poznań lebiega@gmail.com
	Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group
	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
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-24	

Usage

Foreseen use	NLP applications
NLP-specific use	Named entity recognition Machine translation

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

resource type		
Lexical conceptual resource encoding	Encoding level	Semantics Other
	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
Romanian		
Language ID	RO	
Language script	Latin	
Size	19000 entries	
Modality	Modality type	Written language
Size	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 1438 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. The presented resource results from an automatic conversion of the aforementioned dictionary, manually reviewed to include correct information about a number of new features, including sentential subjects, passivisation, and control relations.
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
Organization	Institute of Computer Science, Polish Academy of Sciences Linguistic Engineering Group

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.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	2012-07-20	
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	Web application designed for manual edition of valence frames

Texts

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

Linguality type	Monolingual	
Languages	Polish	
	Language ID	PL
Size	1438 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	http://clip.ipipan.waw.pl/Summarizer	

location		
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	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

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

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

Metadata

Creation date	2012-07-26	
Metadata creators	Marcin Miłkowski	
	Position	Assistant Professor
	Contact	Nowy Świat 72 00-330 Warsaw mmilkows@ifispan.waw.pl http://zil.ipipan.waw.pl/MarcinMiłkowski
	Organization	Institute of Philosophy and Sociology, Polish Academy of Sciences Department of Logic and Cognitive Science
	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
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-25	

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		LanguageTool

	short name	
	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
Creation start date	2006-08-17	

Resource documentation

Reports	Miłkowski, Marcin. 2010. "Developing an Open-source, Rule-based Proofreading Tool." <i>Software: Practice and Experience</i> 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	Segmentation	
	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

Distribution

Availability	Available – unrestricted use	
IPR holder	Ida Kurcz	
	Contact	un@known.pl
	Andrzej Lewicki	
	Contact	un@known.pl
	Jadwiga Sambor	
	Contact	un@known.pl
	Krzysztof Szafran	
	Contact	k.szafran@mimuw.edu.pl
	Jan Paweł 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

	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	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 standoff	True
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 Anotatornia
	Annotation tool	Anotatornia
	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>Głowińska K., Przepiórkowski A. The Design of Syntactic Annotation Levels in the National Corpus of Polish. W: LREC 2010 proceedings.</p> <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>
----------------	---

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

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
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

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
	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
Source	CESAR	
Metadata language ID	en	
Metadata last date updated	2012-07-17	

Usage

Foreseen use	NLP applications	
NLP-specific use	Pos tagging Morphosyntactic tagging	
Actual uses	NLP applications	
	NLP-specific use	Pos tagging Morphosyntactic 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	Pos tagging Morphosyntactic 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	Pos tagging Morphosyntactic 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	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
	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
	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	
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 Morfeusz (libmorfeusz.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

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://ltc.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

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	Attribution Academic - non-commercial use No derivatives
Access medium	Downloadable
Download location	http://lrc.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	Semantic role labelling Document classification

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
Creation start date	2006-01-01	

Resource documentation

Reports	<p>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.</p> <p>Pala, K., Horák, A., Rambousek, A., Vetulani, Z., Konieczka, P., Marciniak, J., 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</p>
----------------	--

	<p>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 – relations Semantics – cross references 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	13200 synsets	

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/resources/parallel/pelcra_par_1.tgz
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
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
---------------------	---

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/resources/parallel/pelcra_par_1.tgz	
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
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	Piotr Pęzik

creators	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
	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
Source	CESAR	
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	75700000 tokens	
Tool	xmllint	
Validator	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
	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
--	---------------------	---

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
	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
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	eng	
	Size	3500000 tokens	
	Polish		
	Language ID	pol	
	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
		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
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
	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
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	eng	
	Size	32400000 tokens	
	Polish		
	Language ID	pol	
	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
	Alignment		
	Segmentation level	Sentence	
	Format	text/xml	

	Conformance to standards best practices	TEI
	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
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	eng
	Size	4200000 tokens
	Polish	
	Language ID	pol
	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
		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
	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	<table border="1"> <tr> <td colspan="2">Piotr Pęzik</td> </tr> <tr> <td>Position</td> <td>assistant professor</td> </tr> <tr> <td>Contact</td> <td>Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl</td> </tr> <tr> <td>Organization</td> <td>University of Łódź PELCRA group, Chair of English Language and Applied Linguistics</td> </tr> <tr> <td colspan="2">Drózdź Łukasz</td> </tr> <tr> <td>Position</td> <td>IT specialist</td> </tr> <tr> <td>Contact</td> <td>Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl</td> </tr> <tr> <td>Organization</td> <td>University of Łódź PELCRA group, Chair of English Language and Applied Linguistics</td> </tr> </table>	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	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
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																	
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																	
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/resources/parallel/pelcra_par_2.tgz

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
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

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/resources/parallel/pelcra_par_2.tgz
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
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
	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
Source	CESAR	
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	710000 tokens	
Tool	xmllint	
Validator	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
	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

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
	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
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 semgent alignment tool (http://kilgray.com/products/memoq).	
Languages	Polish	
	Language ID	pol
	Size	323000 tokens
	English	
	Language ID	eng
	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	MemoQ

tool		
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
	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
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
	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
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/resources/parallel/pelcra_sp_1.tgz
Version	1.0
Revision	compilation of the corpus
Last update	2011-09-30

Contacts

Piotr Pezik	
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

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

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/resources/parallel/pelcra_sp_1.tgz	
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
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
	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	
Source	CESAR	
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	1400000 tokens	
Tool	xmllint	
Validator	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

	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

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
	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
Funding projects	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
	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	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	pol
	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 standards best practices	TEI
	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
---------------------	--

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
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
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
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
Metadata language name	en	
Metadata language ID	en-gb	
Metadata last date updated	2012-06-15	

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

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	CESAR

	name	
	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
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/ecl-dictionaries
Tool documentation type	Online

Lexical conceptual resource

Lexical	Terminological resource
----------------	-------------------------

conceptual resource type		
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:rdf+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

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
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
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
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
Metadata language name	en	
Metadata language ID	en-gb	
Metadata last date updated	2012-06-15	

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

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/pelcra-word-aligned-corpora	
	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
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 LRs	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 only supports binary classification scenarios in which one wants to detect one of two possible languages. A model for distinguishing between Polish and English is provided with the software. The language detector uses the Weka implementation of machine learning classification. In particular, the default language detector provided in the package uses a binary support vector machine classifier implementation.
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 Pezik	
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

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
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
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
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
Metadata language name	en	
Metadata language ID	en-gb	
Metadata last date updated	2012-07-09	

Usage

Resource associated with	http://www.cs.waikato.ac.nz/ml/weka/		
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
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 software	http://www.cs.waikato.ac.nz/ml/weka/	
Required hardware	None	
Required LRs	http://sourceforge.net/projects/morfologik/files/morfologik-stemming/	
Running	JRE	

environment details		
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 15 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
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

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
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
	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
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	2063000 words	
Validation report	Reports	All files are valid XML conforming to the TEI P5 and XLIFF schemas.
Tool	xmllint	
Validator	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
	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
--	---------------------	--

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
	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
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	Parallel

type			
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	1137000 words	
	Polish		
	Language ID	pl	
	Language script	Latn	
	Size	926000 words	
Modality	Modality type	Written language	
	Size	15 texts	
Size	15 texts, 2063000 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	2012-06-30	
	Size	15 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
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
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	2012-06-30	
	Size	15 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
		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
Domains	literature		
Time coverages	1833-1939		
Creation	Original source	http://gutenberg.org http://wikisource.org http://wolnelektury.pl	
	Creation	Manual	

	mode	
	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
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

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
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
	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
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	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
	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

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
	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
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
		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
	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
		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
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 script	Latn
	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
		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
	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
	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
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	Latn	

script	
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
	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
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
	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
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 script	Latn
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
		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	

		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
		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
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 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	509
Resource type	Corpus
URL	http://pelcra.pl/res/parallel/pelcra-par-4
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
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

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	Attribution Academic - non-commercial use	
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
	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
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	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
	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

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
	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
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	Latn	

	script		
	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
		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	
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	<table border="1"> <tr> <td colspan="2">Piotr Pezik</td> </tr> <tr> <td>Position</td> <td>assistant professor</td> </tr> <tr> <td>Contact</td> <td>Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl</td> </tr> <tr> <td>Organization</td> <td>University of Łódź PELCRA group, Department of English Language and Applied Linguistics</td> </tr> <tr> <td colspan="2">Łukasz Drózdź</td> </tr> <tr> <td>Position</td> <td>IT specialist</td> </tr> <tr> <td>Contact</td> <td>Kościuszki 65 90-514 Łódź contact@pelcra.pl http://pelcra.pl</td> </tr> <tr> <td>Organization</td> <td>University of Łódź PELCRA group, Department of English Language and Applied Linguistics</td> </tr> </table>	Piotr Pezik		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	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
Piotr Pezik																		
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																	
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																	
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/resources/parallel/pelcra_sp_2.tgz
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
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

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	Attribution Academic - non-commercial use Other	
Access medium	Downloadable	
Download location	http://pelcra.pl/resources/spoken/pelcra_sp_2.tgz	
Attribution text	Pęzik, P. (2012). Język mówiony w NKJP. In Narodowy Korpus Języka Polskiego. Wydawnictwo Naukowe PWN, Warsaw. Forthcoming.	
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
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

	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
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	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
	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

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
	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
Funding projects	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
	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	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	Speaker noise Background noise
	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 – 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	Noise Music 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 persons	Mixed	
		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 Pezik	
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
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

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
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
	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
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
	Michał Margielewski	
	Position	developer
	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
	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
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/weblign
Tool documentation type	Online

Tool/service

Tool/service type	Tool	
Tool/service subtype	web crawler	
Language dependent	False	
Output	UTF-8	
	Media type	text
	Modality type	Written language
Operating system	OS-independent	
Required software	Java Runtime Environment	
	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

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
Availability start date	2012-06-15	

Licences

CC BY-NC		
Restrictions of use	Attribution Academic - non-commercial use	
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
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
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
Metadata	en	

language name	
Metadata language ID	en-gb
Metadata last date updated	2012-07-04

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

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
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++

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 17,552 synsets (literals 29,565): 1183 adjectives (literals 1557), 2043 verbs (literals 3793), 14,275 nouns (literals 24,147), other 51. 702 synsets are not connected to the PWN, being either Balkan specific concepts (532) or Serbian specific concepts (171). 16,487 synsets have definitions in Serbian, and 1,265 have examples of usage. Semantic relations in SrpWN: hypernym - 16,047; holo_part - 1,275; holo_member - 3,682; holo_portion - 118; near_antonym - 712; be_in_state - 245; causes - 60. From 29,565 literals in SrpWN 9,676 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
Last update	2012-07-22

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

Distribution

Availability	Available – restricted use
IPR holder	Cvetana Krstev
	Position

	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana/
Availability start date	2011-12-01	

Licences

MSCommons_NoCOM-NC-NR		
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	2012-07-27	

Validation

Validated	True	
Type	Formal	
Size	17552 synsets	
Tool	XMLSpy; VisDic	
Validator	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs

Usage

Access tool	VisDic; LeXimir; VeBranka: Bibliša		
Foreseen use	Human use NLP applications		
NLP-specific use	Semantic role labelling Document classification		
Actual uses	NLP applications		
	NLP-specific use	Information extraction Information retrieval Web services	
	Reports	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. 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. 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.	
	Usage project	Serbian Language and its Resources: Theory, Description and Applications	
		Project short name	
		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		
Actual use details			

Resource creation

Resource creator	Gordana Pavlović-Lažetić	
	Position	Professor
	Contact	gordana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~gordana/
	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
	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	
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

Reports	
Samples location	http://
Tool documentation type	Online

Lexical conceptual resource

Lexical conceptual	Wordnet
---------------------------	---------

resource type		
Lexical conceptual resource encoding	Encoding level	Semantics
	Linguistic information	Semantics – relations Part of speech
	Conformance to standards best practices	Word net
	Theoretic model	
	External ref	
	Extratextual information	Images

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	17552 synsets		
Modality	Modality type	Written language		
Size	17552 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 (Pešćanik), 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	vitas@matf.bg.ac.rs http://www.matf.bg.ac.rs/~vitas
Organization	University of Belgrade Faculty of Mathematics

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
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	vitas@matf.bg.ac.rs

		http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics
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

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 Assistent
	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 Assistent
	Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs
	Organization	University of Belgrade Faculty of Philology

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 Morphological analysis Morphosyntactic tagging Lexicon acquisition from corpora	
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 Assistent
	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 Lemmatized 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 Assistent
Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs
Organization	University of Belgrade Faculty of Philology

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
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
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

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 Assistent
	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 Morphological analysis Morphosyntactic tagging Lexicon acquisition from corpora	
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

		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 31 literary texts: 28 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 1,738,752 words (953,935 in the French part, 784,817 in the Serbian part). More about the content of this corpus can be found at: http://www.korpus.matf.bg.ac.rs/SrpFranKor/SrpFranKor_2012_07.pdf
Identifier	604
Resource type	Corpus
URL	http://www.korpus.matf.bg.ac.rs/SrpFranKor
Version	v2.1
Last update	2012-07-25

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

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
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	vas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vas
	Organization	University of Belgrade Faculty of Mathematics
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

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	2012-07-27	

Validation

Validated	True	
Type	Formal	
Mode	Mixed	
Extent	Full	
Size	51 files	
Validator	Duško Vitas	
	Position	Professor
	Contact	vas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vas
	Organization	University of Belgrade Faculty of Mathematics

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	Lexicon acquisition from corpora Bilingual lexicon induction 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	Parallel

type			
Languages	Serbian		
	Language ID	srp	
	Language script	Latin	
	Size	784817 words	
	French		
	Language ID	fra	
	Language script	Latin	
	Size	953935 words	
Size	1738752 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	2012-07-01	
	Annotators	Duško Vitas	
		Position	Professor
Contact		vas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas	
Domains	general (103475 words) literature (1635277 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_12.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

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
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	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Organization	University of Belgrade Faculty of Mathematics
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

Metadata

Creation date	2012-07-08	
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	2011-11-18	

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

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	Lexicon acquisition from corpora Bilingual lexicon induction Temporal expression recognition Terminology extraction
	Reports	Duško Vitas, Svetla Koeva, Cvetana Krstev, Ivan Obradović, “Tour du monde through the dictionaries”, Actes du 27eme Colloque International sur le Lexique et la Gammaire, 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	vas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vas
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	55816 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		vas@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

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
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

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	text audio image video
Output	Media type	text audio image video
Operating system	Linux	
Required software	NXDB eXist	
Required hardware	None	
Required LRs	none	
Tool/service evaluation	Evaluated	True
	Level	Diagnostic
	Evaluators	Ivana Tanasijević
	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 4,420,711 words (2,330,742 in the English part, 2,089,969 in the Serbian part). More about the content of this corpus can be found at: http://www.korpus.matf.bg.ac.rs/SrpEngKor/SrpEngKor_2012_07.pdf
Identifier	607
Resource type	Corpus
URL	http://www.korpus.matf.bg.ac.rs/SrpEngKor
Version	v1.0
Last update	2012-08-01

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

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
Availability start date	2012-08-01	

Licences

CC BY-NC	
Restrictions of use	Academic - non-commercial use
Access medium	Accessible through interface
Execution	http://www.korpus.matf.bg.ac.rs/SrpEngKor

location		
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
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

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	2012-07-27	

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

Usage

Access tool	Corpus query processor (CQP)
Resource	IMS Open Corpus Workbench (CWB), http://cwb.sourceforge.net

associated with		
Foreseen use	Human use NLP applications	
NLP-specific use	Lexicon acquisition from corpora	
Actual uses	Human use	
	NLP-specific use	Lexicon acquisition from corpora Bilingual lexicon induction 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	Duško Vitas	
	Position	Professor
	Contact	vitas@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~vitas
	Miloš Utvić	
	Position	Teaching Assistent
	Contact	misko@matf.bg.ac.rs
	Cvetana Krstev	
	Position	Professor
	Contact	cvetana@matf.bg.ac.rs http://poincare.matf.bg.ac.rs/~cvetana/
	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
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	2089969 words	
	English		
	Language ID	eng	
	Language script	Latin	
	Size	2330742 words	
Size	4420711 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	2012-08-01	
	Annotators	Miloš Utvić	
		Position	Teaching Assistant
Contact		misko@matf.bg.ac.rs	
Domains	general (895091 words)		

	literature (1079154 words) science (371867 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	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).
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

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	Ranka Stanković	
	Position	Assistant professor
	Contact	ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en
	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	2012-07-25	

Validation

Validated	True	
Type	Formal	
Size	85868 entries	
Tool	NooJ;LeXimir	
Validator	Ranka Stanović	
	Position	Assitant Professor
	Contact	ranka@rgf.bg.ac.rs http://www.rgf.bg.ac.rs/profesor.php?id=219&lang=en
	Miloš Utvić	
	Position	Teaching Assistent
	Contact	misko@matf.bg.ac.rs

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	<p>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 Silberztein, pp. 147-168, Presses Universitaires de Franche Comté, Besancon, 2007.</p> <p>Sandra Gucul-Milojević, Vanja Radulović, and Cvetana Krstev. "Usage 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</p>

		2011 International Nooj Conference, pp. 96-108, Cambridge Scholars Publishing, 2012. ISBN (13) 978-1-4438-3711-8.
Usage project	Serbian Language and its Resources: Theory, Description and Applications	
	Project short name	
	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
Actual use details		

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
	Organization	University of Belgrade Faculty of Geology and Mining
	Miloš Utvić	
	Position	Teaching assistant
	Contact	misko@matf.bg.ac.rs
	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	
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	2011-04-01	

Resource documentation

Reports	
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 Part of speech Inflection
	Theoretic model	
	External ref	

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

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

--

MSCommons_NoCOM-NC-NR		
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	Morphosyntactic tagging Lemmatization Morphological analysis		
Actual uses	NLP applications		
	NLP-specific use	Lemmatization Morphological analysis Morphosyntactic tagging	
	Reports	Cvetana Krstevy, 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 short name	
		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		
Actual use details			

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

Reports	
Samples location	http://
Tool documentation type	Online

Lexical conceptual resource

Lexical conceptual	Computational lexicon
---------------------------	-----------------------

resource type		
Lexical conceptual resource encoding	Encoding level	Morphology
	Linguistic information	Lemma Part of speech Inflection
	Conformance to standards best practices	MULTEXT
	Theoretic model	
	External ref	

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

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
Availability start date	2012-07-31	

Licences

CC BY-NC		
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

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

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	Pos tagging Morphological analysis Morphosyntactic tagging	
Actual uses	NLP applications	
	NLP-specific use	Pos tagging
	Reports	Tomaž Erjavec, Cvetana Krstev, Vladimir Petkevič, Kiril Simov, Marko Tadić, Duško Vitas, “The MULTEXT-East

	<p>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, “MULTEXT-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	Cvetana Krstev	
	Position	Assistent 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
	Miloš Utvić	
	Position	Teaching Assistent
	Contact	misko@matf.bg.ac.rs http://www.fil.bg.ac.rs
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

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
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

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	
	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 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

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
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	<p>The Corpus of contemporary Serbian Newspapers and Magazines, SrpNovKor, consists of about 3,3 million 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", "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", "Economist", "Ekonomaska politika", "Ekspres", "Elle", "em na kvadrat", "EM portal", "Enterijer", "ESTETSKA hirurgija & kozmetika", "EU market", "Evropa", "Exclusive", "Fame", "FHM", "FHM collection", "Fito Doktor", "Fly &</p>

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", "Ilustrovana 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", "Officiel", "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".

Identifier	612
Resource type	Corpus

	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

Distribution

Availability	Available – unrestricted use	
IPR holder	Velimir Curgus	
	Position	Director
	Contact	vcurgus@arhiv.rs
	Organization	Ebart consulting d.o.o. Media archive
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
Distribution rights holder	Velimir Curgus	
	Position	Director
	Contact	vcurgus@arhiv.rs
	Organization	Ebart consulting d.o.o. Media archive

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
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
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	Automatic	

	mode	
	Creation mode details	IMB Lotus Notes

7. IBL resources

7.1. Bulgarian National Corpus

General Information

Short name	BulNC
Description	The Bulgarian National Corpus (BulNC) is a large representative publicly available corpus focused on Bulgarian. It 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). The BulNC is compiled mainly for the purposes of computational research and implementations. BulNC is constantly enlarged and developed. The aim is to ensure representativeness and balance of the data by including texts from different modality (written and spoken), various time periods, domains and genres. BulNC has been substantially expanded and it now contains 979.6 million tokens (during the last six months it has been increased with app. 50%). Currently, written texts comprise 91.11% of the corpus while spoken texts represent 8.89%. Three basic approaches are implemented for collecting new samples for BulNC: use of readily developed collections of texts; manual collection (by means of Internet browsing and downloading texts); and automatic collection (by means of web crawling). The texts in the corpus are published between 1878 and 2011, and the majority - after 2000. All texts are supplied with extensive metadata description compliant with the well established standards. The metadata comprise of 25 fields. Each text is supplied with editorial metadata (author's name, text title, source, etc.) and classificatory metadata (general category, domain, genre). The values in the classificatory information columns are limited to a list of predetermined options which ensures a harmonised approach towards the description of the samples. 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), etc.
Identifier	801
Resource type	Corpus
URL	http://ibl.bas.bg/en/BGNC_en.htm
Version	4.0
Last update	2012-07-20

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 and Department of Lexicology and Lexicography
	Contact 52 Shipchenski prohod Blvd., Bl. 17 1113 Sofia bulnc@dcl.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

Metadata

Creation date	2011-11-20
----------------------	------------

Validation

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

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use
	NLP applications

Resource creation

Funding projects	Bulgarian National Corpus project	
	Funding type	National funds

	Central and South-East European Resources	
	Funding type	EU funds

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
Size	979600000 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
	Morphosyntactic annotation – POS tagging	
	Segmentation level	Word
	Semantic annotation – word senses	
	Segmentation level	Word

7.2. Bulgarian National Corpus Collocation service

General Information

Short name	BulNColl
Description	The Corpus Collocation service gives access to the Bulgarian National Corpus. The Bulgarian National Corpus (BulNC) is a large representative corpus of Bulgarian, publicly available for online queries for concordances and collocations. As of November 2011 the core of BulNC includes 144,669 text samples which comprise of 470.1 million tokens. A uniform framework was developed for structuring BulNC, data storage format and description of

	<p>the texts. Each text is supplied with editorial metadata (author's name, text title, source, etc.) and classificatory metadata (general category, domain, genre). 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), etc. The Corpus Collocation 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=netuser: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. Collocations have numerous applications: corpus linguistics and computational linguistics (tasks for machine translation, text generation, summary generation and others). The users may observe the frequency of words and language constructions, and generate frequency lists and language models.</p>
Identifier	802
Resource type	Tool/service
Tool/service type	Service
URL	http://ibl.bas.bg/en/BGNC_en.htm
Version	1.0
Last update	2011-11-20

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 bulnc@dcl.bas.bg http://ibl.bas.bg/en/BGNC_en.htm
Availability start date	2011-11-20	

Licences

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
----------------------	------------

Validation

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

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use NLP applications

Resource creation

Funding projects	Bulgarian National Corpus project	
	Funding type	National funds
	Central and South-East European Resources	
	Funding type	EU funds

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
		Contact	boby@dcl.bas.bg http://dcl.bas.bg/en/people_en/boby.html

7.3. Bulgarian Part-of-Speech Corpus

General Information

Short name	BulPosCor
Description	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.
Identifier	803
Resource type	Corpus
URL	http://dcl.bas.bg/poscor/BulPosCor.html
Version	1.0
Last update	2011-11-20

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 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
----------------------	------------

Validation

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

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use NLP applications

Resource creation

Funding projects	Bulgarian National Corpus project	
	Funding type	National funds
	Central and South-East European Resources	
	Funding type	EU funds

Texts

Media type	text	
Linguality type	Monolingual	
Languages	Bulgarian	
	Language ID	bg
Size	217000 tokens	
Character encoding	UTF-8	
Annotation	Segmentation	
	Segmentation level	Paragraph
	Segmentation	
	Segmentation	Sentence

level	
Segmentation	
Segmentation level	Word
Lemmatization	
Segmentation level	Word
Morphosyntactic annotation – below POS tagging	
Segmentation level	Word
Morphosyntactic annotation – POS tagging	
Segmentation level	Word

7.4. 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/BulSemCor.html
Version	3.0
Last update	2011-11-20

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 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
----------------------	------------

Validation

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

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use
	NLP applications

Texts

Media type	text
Linguality type	Monolingual

Languages	Bulgarian	
	Language ID	bg
Size	99000 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
	Morphosyntactic annotation – POS tagging	
	Segmentation level	Word
	Semantic annotation – word senses	
	Segmentation level	Word

7.5. 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 33 languages – English, German, French, Slavic and Balkan languages, as well as other European and non-European languages (28 languages are available through the web interface). 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. At present, the Bulgarian-X Language</p>

	Parallel Corpus contains 1.9 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 5 other corpora between 100 and 200 million tokens per language, 16 parallel corpora of size in the range 30-52 million tokens per language, further 7 in the range 1-10 million tokens, and the rest are below 1 million, with the smallest corpora being the Chinese, Japanese and Icelandic with less than 50,000 tokens per language. 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.
Identifier	805
Resource type	Corpus
URL	http://dcl.bas.bg/bulXcor/BulXCor.html
Version	2.0
Last update	2012-07-20

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 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

Metadata

Creation date	2011-11-20
----------------------	------------

Validation

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

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use

Texts

Media type	text
Linguality type	Multilingual
Multilinguality type	Parallel
Languages	Bulgarian
	Language ID bg
	English
	Language ID en
	Romanian
	Language ID ro
	Greek
	Language ID el
	Czech
	Language ID cs
	Polish
	Language ID pl
	Slovak
	Language ID sk
	Spanish
	Language ID es
	Danish
	Language ID da

Finish	
Language ID	fi
Hungarian	
Language ID	hu
Estonian	
Language ID	et
Slovenian	
Language ID	sl
German	
Language ID	de
Lithuanian	
Language ID	lt
Italian	
Language ID	it
Bosnian	
Language ID	bs
Galician	
Language ID	ga
Croatian	
Language ID	hr
Latvian	
Language ID	lv
Macedonian	
Language ID	mk
Maltese	
Language ID	mt
Dutch	
Language ID	nl
Portuguese	
Language ID	pt
Albanian	
Language ID	sq
Swedish	
Language ID	sv
Turkish	
Language ID	tr
Size	1900000000 tokens
Character encoding	UTF-8
Annotation	Alignment
	Segmentation

	level	
	Segmentation	
	Segmentation level	Sentence
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation – below POS tagging	
	Segmentation level	Word

7.6. Bulgarian WordNet

General Information

Short name	BulNet
Description	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 BlakaNet project. Bulgarian WordNet describes meaning of a lexical unit by placing it within a network of semantic relations, such as hypernymy, meronymy, antonymy etc. The Bulgarian wordnet is one of the most complete and consistent lexical resources (in comparison the literals in the Bulgarian wordnet are much greater in number than the word list in a standard spelling dictionary). The synonym sets from different languages are connected by means of inter-language equivalence relations, which are used as a basis for the development of the wordnet multilingual lexical-semantic network, the so called global wordnet. The Bulgarian wordnet is approximately one quarter the size of the English wordnet and is one of the biggest in Europe
Identifier	806
Resource type	Lexical conceptual resource
Lexical conceptual resource type	Wordnet
URL	http://catalog.elra.info/product_info.php?products_id=802
Version	4.0
Last update	2012-02-20

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

Restrictions of use	Other
Access medium	CD-ROM

Metadata

Creation date	2011-11-27
----------------------	------------

Validation

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

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use
	NLP applications

Resource creation

Funding projects	BalkaNet	
	Funding type	EU funds
	BulNet - A Lexical-Semantic Network of Bulgarian	
	Funding type	National funds
	Central and South-East European Resources	
	Funding type	EU funds

Creation start date	2001-09-01
----------------------------	------------

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>
----------------	---

Lexical conceptual resource

Lexical conceptual resource type	Wordnet	
Lexical conceptual resource encoding	Encoding level	Semantics
	Linguistic information	Part of speech Semantics – relations Semantics – cross references 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	Cyrilic
Modality	Modality type	Written language
Size	41668 synsets	
Character encoding	UTF-8	

7.7. WordNet web service

General Information

Short name	WordNetWeb
Description	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 BlakaNet project. Wordnet service is an online service that gives the users access to a subset of the Bulgarian wordnet (BulNet), containing over 1200 synonym sets (synsets) from the so called Base Concepts sunset 1 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 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 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

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	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
----------------------	------------

Validation

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

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use

Resource creation

Funding projects	BalkaNet	
	Funding type	EU funds
	BulNet - A Lexical-Semantic Network of Bulgarian	
	Funding type	National funds
	Central and South-East European Resources	
Funding type	EU funds	
Creation start date	2001-09-01	

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, Tadič et al. (eds) Proceedings of the 6th International Conference on Formal Approaches to</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
	Contact	boby@dcl.bas.bg http://dcl.bas.bg/en/people_en/boby.html	

7.8. 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 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. The table below shows the operational compatibility of WinEst with the various versions of Microsoft</p>

	Windows and Microsoft Office.
Identifier	808
Resource type	Tool/service
Tool/service type	Tool
URL	http://dcl.bas.bg/est/index_en.php
Version	2.0
Last update	2011-11-20

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 est@dcl.bas.bg http://dcl.bas.bg
Availability start date	2011-06-01	

Licences

Restrictions of use	Other
Access medium	Downloadable
Download location	http://dcl.bas.bg/sites/default/files/webfm/WinEst/winestSetup.exe

Metadata

Creation date	2011-11-20
----------------------	------------

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	
	Funding type	EU funds
Creation start date	2010-01-01	

Resource documentation

Reports	Oliva K. i Sv. Koeva – Sintaksis na nevezmozhnoto, Balgarski ezik, 3, 7-17, 2009. ISSN 0005-4283. ERIH
----------------	--

Tool/service

Tool/service type	Tool		
Language dependent	True		
Input	Media type	text	
Output	Media type	text	
Operating system	Windows		
Tool/service evaluation	Evaluated	True	
	Level	Diagnostic	
	Evaluators	Angel Genov	
		Position	Assistant
	Contact	angel@dcl.bas.bg http://dcl.bas.bg/PersonalPages/angel/	

7.9. 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

	<p>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 intergartion) 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. WenEst 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.</p>
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

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 est@dcl.bas.bg http://dcl.bas.bg
Availability start date	2011-06-01	

Licences

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
----------------------	------------

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	
	Funding type	EU funds
Creation start date	2010-01-01	

Resource documentation

Reports	Oliva K. i Sv. Koeva – Sintaksis na ne vazmozhnoto, Balgarski ezik, 3, 7-17, 2009. ISSN 0005-4283. ERIH
----------------	---

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	Angel Genov	
		Position	Assistant
	Contact	angel@dcl.bas.bg http://dcl.bas.bg/PersonalPages/angel/	

7.10. Bulgarian-X Language Parallel Corpus Collocation service

General Information

Short name	BulXLColl
Description	Collocations service is a web service for collocations search and different types of statistics over the Bulgarian-X Language Parallel Corpus. Bulgarian-X Language Parallel Corpus includes parallel corpora of 33 languages – English, German, French, Slavic and Balkan languages, as well as other European and non-European languages (28 languages are available through the web interface). At present, the Bulgarian-X Language Parallel Corpus contains 1.9 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 5 other corpora between 100 and 200 million tokens per language, 16 parallel corpora of size in the range 30-52 million tokens per language, further 7 in the range 1-10 million tokens, and the rest are below 1 million, with the smallest corpora being the Chinese, Japanese and Icelandic with less than 50,000 tokens per language. Each parallel subcorpus within Bul-X-Cor mirrors the structure of BulNC. The Corpus Collocation 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=нетuser: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=нет&lang=bg
Identifier	810
Resource type	Tool/service
Tool/service type	Service
URL	http://ibl.bas.bg/en/BGNC_en.htm
Version	1.0
Last update	2012-07-20

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 bulnc@dcl.bas.bg http://ibl.bas.bg/en/BGNC_en.htm
Availability start date	2011-11-20	

Licences

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	2012-07-20
----------------------	------------

Validation

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

Usage

Foreseen use	Human use NLP applications
Actual uses	Human use
	NLP applications

Resource creation

Funding projects	Bulgarian National Corpus project	
	Funding type	National funds
	Central and South-East European Resources	
	Funding type	EU funds

Tool/service

Tool/service type	Service
Language	False

dependent			
Input	Media type	text	
Output	Media type	text	
Operating system	OS-independent		
Tool/service evaluation	Evaluated	True	
	Level	Diagnostic	
	Evaluators	Borislav Rizov	
		Position	Assistant
Contact		boby@dcl.bas.bg http://dcl.bas.bg/en/people_en/boby.html	

7.11. 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

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 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
----------------------	------------

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

Funding projects	Central and South-East European Resources	
	Funding type	EU funds
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	

7.12. Bulgarian Frequency Dictionary

General Information

Short name	BulFreq
Description	Bulgarian Frequency Dictionaries are lemma frequency dictionaries extracted from the Bulgarian National Corpus (BulNC) which was annotated at various linguistic levels - sentence segmentation, POS tagging,

	lemmatisation, etc. BulNC contains 6 domain-specific subcorpora and thus a 6 domain-specific Frequency Dictionary were developed independently, as well as a general dictionary which combines all domain-specific ones. Each dictionary is in 2 variants: in alphabetical order and in frequency order. The frequencies are automatically collected and more efficient methods for compilation of frequency lists and dictionaries are still being investigated. The compilation of the 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/dictionaries_en.html
Version	1.0
Last update	2012-07-20

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 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/Frequency/Frequency.zip
Fee	free of charge

Metadata

Creation date	2012-07-27
----------------------	------------

Validation

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

Usage

Foreseen use	Human use NLP applications
NLP-specific use	Named entity recognition
Actual uses	Human use
	NLP applications

Resource creation

Funding projects	Central and South-East European Resources	
	Funding type	EU funds
Creation start date	2012-01-01	

Resource documentation

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

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	2142555 words
Character encoding	UTF-8

7.13. 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/programs_en.html
Version	3.0
Last update	2012-07-20

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

		est@dcl.bas.bg http://dcl.bas.bg
Availability start date	2005-06-01	

Licences

LGPLv3		
Restrictions of use	Share alike	
Access medium	Downloadable	
Download location	http://dcl.bas.bg/Tools/Hydra/Hydra.zip	

Metadata

Creation date	2012-07-20
----------------------	------------

Validation

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

Usage

Foreseen use	Human use
Actual uses	Human use

Resource creation

Funding projects	Central and South-East European Resources	
	Funding type	EU funds
Creation start date	2010-01-01	

Resource documentation

Reports	Rizov, B. Hydra: A Modal Logic Tool for Wordnet Development, Validation and Exploration, 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
----------------	---

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	Rizov Borislav	
		Position	Assistant
	Contact	boby@dcl.bas.bg http://dcl.bas.bg/en/people_en/boby.html	

7.14. Chooser - annotation tool

General Information

Short name	Chooser
Description	<p>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.</p>
Identifier	814
Resource type	Tool/service
Tool/service type	Tool
URL	http://dcl.bas.bg/en/programs_en.html
Version	3.0
Last update	2012-07-20

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 est@dcl.bas.bg http://dcl.bas.bg
Availability start date	2003-06-01	

Licences

LGPLv3	
Restrictions of use	Share alike
Access medium	Downloadable
Download location	http://dcl.bas.bg/Tools/Chooser/Chooser.zip

Metadata

Creation date	2012-07-20
----------------------	------------

Validation

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

Usage

Foreseen use	Human use
Actual uses	Human use

Resource creation

Funding projects	Central and South-East European Resources	
	Funding type	EU funds
Creation start date	2010-01-01	

Resource documentation

Reports	Koeva, Sv., B. Rizov, S. 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, pp. 728-734, 2008. ISBN 2-9517408-4-0
----------------	--

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	Rizov Borislav
	Position	Assistant
	Contact	boby@dcl.bas.bg http://dcl.bas.bg/en/people_en/boby.html

7.15. 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
Version	3.0
Last update	2012-07-20

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 est@dcl.bas.bg http://dcl.bas.bg
Availability start date	2005-06-01	

Licences

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
----------------------	------------

Validation

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

Usage

--	--

Foreseen use	NLP applications
Actual uses	NLP applications

Resource creation

Funding projects	Central and South-East European Resources	
	Funding type	EU funds
Creation start date	2010-01-01	

Resource documentation

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

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/	

7.16. Web based infrastructure for Bulgarian data processing

General Information

Short name	DCLservices
Description	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

	<p>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 ofa 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/dclservices/registration/
Version	1.0
Last update	2012-07-20

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 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
Execution location	http://dcl.bas.bg/dclservices/registration/

Metadata

Creation date	2012-07-20
----------------------	------------

Validation

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

Usage

Foreseen use	NLP applications
Actual uses	NLP applications

Resource creation

Funding projects	Central and South-East European Resources	
	Funding type	EU funds
Creation start	2010-01-01	

date	
------	--

Resource documentation

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

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. LSIL resources

8.1. Slovak National Corpus

General Information

Short name	prim
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/
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	Eudovít Štúr Institute of Linguistics	
	Short name	JÚEŠ
	Department name	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	2012-07-16	

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

8.2. 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	Ľudovít Štúr Institute of Linguistics	
	Short name	JÚĽŠ
	Department name	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 Garabik	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2012-07-16	

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	Word

	level	
	Speech annotation – phonetic transcription	
	Segmentation level	Word
	Speech annotation – sound events	
	Speech annotation – sound to text alignment	
	Segmentation level	Utterance
	Speech annotation – speaker identification	
	Speech annotation – speaker turns	

8.3. 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	Eudovít Štúr Institute of Linguistics	
	Short name	JÚLŠ
	Department name	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	2011-11-22

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

8.4. Slovak-Czech Parallel Corpus

General Information

Description	Parallel Slovak-Czech corpus is a corpus of sentence aligned texts, mostly
--------------------	--

	fiction. 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	2011-05-17
Last update	2011-05-17

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Eudovít Štúr Institute of Linguistics	
	Short name	JÚĽŠ
	Department name	Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2010-05-09	

Licences

Proprietary	
Restrictions of use	Academic - non-commercial use
Access medium	Accessible through interface
Execution location	http://korpus.juls.savba.sk:8088/

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	2011-11-22	

Texts

Media type	text	
Linguality type	Bilingual	
Multilinguality type	Parallel	
Languages	Slovak	
	Language ID	sk
	Czech	
	Language ID	cs
Size	730000 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

8.5. Slovak-English Parallel Corpus

General Information

Description	The corpus consists of parallel Slovak and English texts (mostly fiction), with automatic lemmatization, morphological analysis (for Slovak), POS tagging (for English). The corpus consists of original English language books and their Slovak translations. 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	2011-08-03
Last update	2011-08-03

Contacts

Radovan Garabík	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Ľudovít Štúr Institute of Linguistics	
	Short name	JÚĽŠ
	Department name	Slovak National Corpus
	Contact	Panská 26 81364 Bratislava korpus@korpus.juls.savba.sk http://korpus.juls.savba.sk/
Availability start date	2011-08-03	

Licences

Proprietary	
Restrictions of use	Academic - non-commercial use
Access medium	Accessible through interface
Execution location	http://korpus.juls.savba.sk:8091/

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	2011-11-22

Texts

Media type	text
Linguality type	Bilingual
Multilinguality type	Parallel
Languages	Slovak
	Language ID sk
	English
	Language ID en
Size	1500000 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

8.6. 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	Ľudovít Štúr Institute of Linguistics	
	Short name	JÚĽŠ
	Department name	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	2012-07-16	

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	

8.7. Balanced Slovak Corpus

General Information

Short name	VYV
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/
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	Eudovít Štúr Institute of Linguistics	
	Short name	JÚĽŠ
	Department name	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	2012-07-16	

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

8.8. 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/
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	Eudovít Štúr Institute of Linguistics	
	Short name	JÚĽŠ
	Department name	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	2012-07-16	

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

8.9. 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 the IRSTLM 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	Eudovít Štúr Institute of Linguistics

	Short name	JÚĽŠ
	Department name	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	2012-07-16	

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

8.10. 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	Ľudovít Štúr Institute of Linguistics	
	Short name	JÚĽŠ
	Department name	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	2012-07-16	

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

8.11. 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	Eudovít Štúr Institute of Linguistics	
	Short name	JÚLEŠ
	Department name	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	2012-07-16	

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

8.12. 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/
Version	1.0
Last update	2011-07-14

Contacts

Radovan Garabik	
Position	researcher
Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk

Distribution

Availability	Available – restricted use	
IPR holder	Eudovít Štúr Institute of Linguistics	
	Short name	JÚĽŠ
	Department name	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 Garabik	
	Position	researcher
	Contact	Panská 26 81364 Bratislava radovan.garabik@kassiopeia.juls.savba.sk
Metadata language ID	en	
Metadata last date updated	2012-07-16	

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	

8.13. 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/
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	Eudovít Štúr Institute of Linguistics	
	Short name	JÚĽŠ
	Department name	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	2012-07-16	

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	Sentence

	level	
	Segmentation	
	Segmentation level	Word
	Lemmatization	
	Segmentation level	Word
	Morphosyntactic annotation – below POS tagging	
	Segmentation level	Word