CESAR

Central and South-East European Resources

Project no. 271022

Deliverable D2.3c
Report on resources (actually or potentially) available to the consortium

Version No. 1.3

31/07/2012

Document Information

Deliverable number:	D2.3c
Deliverable title:	Report on resources (actually or potentially) available to the consortium (name)
Due date of deliverable:	31/07/2012
Actual submission date of deliverable:	31/07/2012
Main Author(s):	Svetla Koeva (IBL)
Participants:	Tamas Varadi (HASRIL) Tibor Pinter (HASRIL) Szaszák György (BME-TMIT) Radovam Garabik (LSIL) Maciej Ogrodniczuk (IPIPAN) Adam Przepiórkowski (IPIPAN) Piotr Pezik (ULodz) Marko Tadic (FFZG) Dusko Vitas (UBG) Cvetana Krstev (UBG) Tsvetana Dimitrova (IBL)
Internal reviewer:	Tamas Varadi (HASRIL)
Workpackage:	2
Workpackage title:	Analysis and selection of language resources
Workpackage leader:	IBL
Dissemination Level:	Public
Version:	1.3
Keywords:	language resources, tools for natural language processing, language technologies

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
1.3	31/07/2012	Final	Tamás Váradi	supervision	
1.2	30/07/2012	draft	Tibor Pinter	editing of the text	
1.1	30/07/2012	draft	Svetla Koeva	Drafting report	

EXECUTIVE SUMMARY
The deliverable gives a detailed description on the actually or potentially available resources to the consortium after M18. The first section provides an in-depth analysis on the criteria of such resources, while the second section summarises the language resources (language by language) gathered in the third six-month period of the project. A more detailed description of the resources is given in the annex.

Table of Contents

1. Background 5
1.1. Project objectives 5
1.2. Baseline situation 5
1.3. Target resources and users 5

2. A common and shared resource description 7
2.1. The metadata scheme developed within T4ME/META-NET 7
2.2. Project specific additions to the scheme 10

3. Resources identified via CESAR between M12 and M18 14
3.1. Summary of the language resources developed in Bulgaria and potentially available to the language engineering community 14
3.2. Summary of the language resources developed in Croatia and potentially available to the language engineering community 16
3.3. Summary of the language resources developed in Hungary and potentially available to the language engineering community 19
3.4. Summary of the language resources developed in Poland and potentially available to the language engineering community 23
3.5. Summary of the language resources developed in Serbia and potentially available to the language engineering community 29
3.6. Summary of the language resources developed in Slovakia and potentially available to the language engineering community 31

4. Conclusions 33

5. Annex 34
5.1. Bulgarian language resources detailed specification 34
5.2. Croatian language resources detailed specification 42
5.3. Hungarian language resources detailed specification 48
5.4. Polish language resources detailed specification 63
5.5. Serbian language resources detailed specification 81
5.6. Slovak language resources detailed specification 89

1. Background

1.1. Project objectives

The CESAR project, in close harmony with the META-NET and sensitive to the dynamics of the community practices, addresses the needs of human language technologies (crucially depending on language resources and tools) by means of enhancing, upgrading, standardizing, and cross-linking of wide variety of language resources and tools, as well as making them accessible to contribute to the development of an open linguistic infrastructure.

The main goals of the CESAR project are:

- to provide a description of the national (resp. language community) landscape in terms of language use; language-savvy products and services, language technologies and resources; main actors (research, industry, government and society); public policies and programmes; prevailing standards and practices; current level of development;
- to contribute to a pan-European digital resource exchange facility by collecting resources and by documenting, linking and upgrading them to agreed standards and guidelines;
- to help build and operate broad, non-commercial, community-driven, interconnected repositories, exchanges, facilities, etc. that can be used by language researchers, developers and professionals;
- to stimulate actions by national and regional actors, public bodies and funding agencies by raising awareness, organizing meetings and other focused events;
- to bridge the technological gap between this region and other parts of Europe by filling obvious and important gaps in language resources and tools infrastructure.

1.2. Baseline situation

The CESAR project is specifically focused on the assembly of basic language resources for six Central and South-East European languages, all of them considered, by any comparison, less-resourced: four of them (Hungarian, Polish, Bulgarian, Slovak) are the official languages of recently acceded EU member states, while two (Croatian and Serbian) are languages of states scheduled to join the EU in the near future. The coverage of these languages brings about an added benefit of the project, anticipating and meeting foreseeable requirements with respect to resources developed for these languages. Building on a wide range of already existing resources and previous national and international activities, the project creates, enhances and operates a comprehensive language-resource platform enabling and supporting large-scale multi- and cross-lingual products and services. In extensive cooperation with the META-NET, resources are upgraded and updated to widely acknowledged standards, thus ensuring interoperability and developing the ground for widespread and efficient potential to modularize them in language technology pipelines.

In the frame of these tasks, language resources and tools already developed or still under development are identified. The *D2.3c Report on resources (actually or potentially) available to the consortium* includes the resources for Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak, identified between the months twelve and eighteenth.

1.3. Target resources and users

The CESAR encompasses a large variety of language resources, including language data, such as written and spoken corpora (annotated and raw, monolingual and multilingual), lexical and terminological databases, grammars, ontologies, etc.; language processing and annotation tools and technologies.

The target users are developers and researchers in both industry and academia. These are private and public institutions, companies and individuals involved in HLT research and development: industrial organizations and SMEs, academic institutions, research organizations, universities, individual researchers and students, national governments, EU institutions, and private investors.

2. A common and shared resource description

The CESAR supports the goal of a common and shared resource description between the four projects constituting METANET (namely: CESAR, METANET4U, META-NORD, and T4ME). The focus was to gather all relevant information (metadata) of the resources actually (or potentially) available. The metadata covers features of resource localization, information about IPR holders (the name of the holder and address of the contact persons), the distribution media (the specified format used for the delivery of the resource), as well as the licence issues and restrictions of its use. The metadata also describe the NLP focused use of the resources in both its actual and its upcoming state (actual and foreseen use). The metadata contain wider information of the resources by offering further readings and publications on the resource, as well as links to the main documentation. The metadata scheme also informs about data types as the media type of the resource or the language covered by the resource.

2.1. The metadata scheme developed within T4ME/META-NET

The CESAR adopted the metadata scheme developed within the T4ME/META-NET project, thus a common metadata description for language resources for many different European languages is provided. Table 1 below describes the metadata scheme with definitions and recommended values used in the T4ME and shared by other three projects that are also part of the META-NET alliance.

	Definition	Recommended Values
resourceTitle	The title is the complete title of the resource without any abbreviations	
resourceName	A short name (e.g., acronym, abbreviation) to identify the language resource.	
IPRholder.organizationShortName		
contact.Person.surname	Surname of the contact person (anyone who can give further information on the resource); when there are more than one contact person, repeat the relevant columns	
contact.Person.givenName	Given name of the contact person (anyone who can give further information on the resource)	
contact.Person.email	Email of the contact person	
availability	Terms of availability; please choose one of the recommended values; if restricted, please specify in restrictionsOfUse	Terms of availability; please choose one of the recommended values; if restricted, please specify in restrictionsOfUse
license	A description of the licensing condition under which the resource can be used; see recommended values for examples	Name of licence, e.g. CC Zero, CC-BY, etc. MSC (IF FOR META-SHARE ONLY). ELRA, LDC, GPL, etc.

distributionMedium	Specify the format used for the delivery of the resource; if possible, use one of the recommended values	internetBrowsing; download; CD-ROM; DVD-R; bluRay; hardDisk; paperCopy; other
restrictionsOfUse	restrictions of use; see recommended values for examples	academic-nonCommercialUse; noDerivatives; shareAlike; attribution; commercialUse (specify details); evaluationUse (specify details if needed); other
licenseSignatory.Person.position	The position (director/head of dept/researcher/etc) of the person in your organisation authorised to sign the licence by which you make the resource available.	
ForeseenUse.foreseenUse	The use for which the resource has been produced. When more than one values use ";" in between	human use; NlpApplications
ForeseenUse.useNLPspecific	the application for which it has been constructed; for indicative values, see recommended values. When more than one values use ";" in between	speech analysis; Discourse analysis; Language identification; Speaker identification; Speaker verification; Speech recognition; Spoken dialogue systems; Voice control; Speech synthesis; Used in project; Face verification; Speech verification; User authentication; Face recognition; Automatic speech recognition; Automatic person recognition; Talking head synthesis; Avatar synthesis; Multimedia development; Voice control; Speech assisted video control; Information retrieval; Word sense disambiguation; Machine Translation; Named Entity recognition; Question answering; Automatic text generation and summarization; Document classification; Emotion recognition; Sign language recognition
ActualUse.actualUse	the actual use of the resource in the framework of a specific project or application	human use; NlpApplications

ActualUse.useNLPspecific	the application in which it has been used; for indicative values, see recommended values. When more than one values use ";" in between	speech analysis; Discourse analysis; Language identification; Speaker identification; Speaker verification; Speech recognition; Spoken dialogue systems; Voice control; Speech synthesis; Used in project; Face verification; Speech verification; User authentication; Face recognition; Automatic speech recognition; Automatic person recognition; Talking head synthesis; Avatar synthesis; Multimedia development; Voice control; Speech assisted video control; Information retrieval; Word sense disambiguation; Machine Translation; Named Entity recognition; Question answering; Automatic text generation and summarization; Document classification; Emotion recognition; Sign language recognition
Description	Description of the resource in prose	

resourceType	type of the resource; please use one of the recommended values	corpus;
		lexicalConceptualResource;
		languageDescription;
		technologyToolService

mediaType	Specification of the media type of the resource; can be multiple if the resource is a multimodal set; please, use one or more of the recommended values	text; audio; video; image; tactile
noLanguages	An indication of the number of languages that are included in the resource.	if one language, then corpus is monolingual
multilingualityType	Whether the corpus is parallel or comparable.	parallel; comparable
languageId	Identifier of the language as defined by ISO 639 that is included in the resource or supported by the tool/service. When more than one value, use ";" in between	ISO 639-3
size	The size of the resource with regard to the SizeUnit measurement in form of a number.	
sizeUnit	Specification of the unit of size that is used when specifying the size; if possible, use one of the recommended values.	word; token; byte; sentence; text; …
annotationType	Specification of the types of annotation levels (tiers) provided by the resource; if possible use recommended values; can be repeated if the values are multiple.	

Table 1. Metadata scheme

2.2. Project specific additions to the scheme

In addition, some new metadata fields are accepted for the metadata scheme developed within the CESAR project. These are as given in Table 2:

	Definition	Recommended Values
projectPartner	The acronym of the partner responsible for collecting the resource.	
resourceLocation	Actual or anticipated location.	
urlDownload	Where to download the resource.	
urlDocumentation	Where information about the resource is published	
resourceSubType	Classification according to the categories used in the resource evaluation for the language whitepaper	Tokenization, Morphology; Parsing; Sentence Semantics; Text Semantics; Advanced Discourse Processing; Information Retrieval; Information Extraction; Language Generation; Summarization, Question Answering, Advanced Information Access Technologies; Machine Translation; Speech Recognition; Speech Synthesis; Dialogue Management; Reference Corpora; Syntax-Corpora; Semantics-Corpora; Discourse-Corpora; Parallel Corpora, Translation Memories; Speech-Corpora; Multimedia and multimodal data; Language Models; Lexicons, Terminologies; Grammars; Thesauri, WordNets; Ontological Resources for World Knowledge; Other

Table 2. Additions accepted in the CESAR project

2.3. Adaptation to the META-SHARE specifications

The specifications used for the description of the language resources at the third D2.3 deliverable are adapted to the common META-SHARE specifications available so far (see Table 3.). The goal is to unify the description of language resources as well as to provide the most important information about them.

resourceName	
resourceShortName	
downloadLocation	if applicable
dateCreation	
projectPartner	
iprHolder.organizationName	
contact.Person.surname	

contact.Person.givenName	
contact.Person.email	
DistributionInfo	please, choose one of the values available-unrestricted use available-restricted use notAvailable underNegotiation
license	
resourceLocation	
distributionAccessMedium	please, leave the appropriate accessibleThroughInterface webExecutable other paperCopy hardDisk bluRay DVD-R CD-ROM downloadable other
restrictionsOfUse	please, leave the appropriate other noModifications informResourceOwner redeposit onlyMSmembers academic-nonCommercialUse evaluationUse commercialUse attribution shareAlike noDerivatives
licenseSignatory.Person.position	
foreseenUse	please, leave the appropriate human use NlpApplications
actualUse	please, leave the appropriate human use NlpApplications
description	
relevantPublications	
resourceType	please, leave the appropriate corpus lexical / conceptual resource language description technology tool / service evaluation package

mediaType	please, leave the appropriate text audio video image sensorimotor
lingualityType	please, leave the appropriate monolingual bilingual multilingual
languageId	
size	
sizeUnit	please, leave the appropriate terms entries turns utterances articles files items seconds elements units minutes hours texts sentences bytes tokens words keywords idiomaticExpressions neologisms multiWordUnits expressions synsets classes concepts lexicalTypes phoneticUnits syntacticUnits semanticUnits predicates phonemes diphones T-HPairs syllables rules other

Table 3. Adaptation to the most recent META-SHARE specifications

There are a number of differently specified descriptions, listed below:

resourceName	vs.	resourceTitle
resourceShortName	vs.	resourceName
downloadLocation	vs.	urlDownload
iprHolder.organizationName	vs.	IPRholder.organizationShortName
DistributionInfo	vs.	availability
distributionAccessMedium	vs.	distributionMedium
lingualityType	vs.	multilingualityType

To focus on the most important information, some specifications are omitted, namely *foreseenUse.useNLPspecific; actualUse.useNLPspecific; urlDocumentation*; *Resource-Subtype; noLanguages; and annotationType*. They will be provided only when resource becomes available through META-SHARE.

3. Resources identified via CESAR between M12 and M18

The *D2.3c Report on resources (actually or potentially) available to the consortium* gives an overview of the main language resources developed in Central-East Europe. It is compiled to give extensive information about the resources for the six languages involved. A table with values of the commonly accepted metadata scheme was constructed through a survey on national level, assisted by national research institutions and private companies, to gather all important information concerning available and potential language resources. As a result of the survey, the description of the resources was made, along with a catalogue of written and spoken language resources to enhance the project actions.

The description gives a detailed view of the main language resources available for the languages covered by the project partners. The description contains language resources for Bulgarian, Hungarian, Croatian, Polish, Serbian, and Slovak languages. The focus was to gather all relevant information (metadata) of the actually (or potentially) available resources.

3.1. Summary of the language resources developed in Bulgaria and potentially available to the language engineering community

The basic resources developed in Bulgaria, many of which are constantly updated, can be classified in the following categories:

- **Multilingual Text Corpora**
    - **Bulgarian-English clause aligned corpus** consists of 363,402 tokens altogether (174,790 for Bulgarian and 188,612 for English) distributed over five thematic domains: Fiction (21.4%), News (37.1%), Administrative (20.5%), Science (11.2%) and Subtitles (9.8%). Both Bulgarian and English parts of the corpus are first automatically segmented and then aligned at sentence level.
- **Lexical Conceptual Resources**
    - **Lists of Bulgarian Multiword Expressions** is a set of 13 lists comprising 27,784 entries (MWEs and phrases), including non-decomposable MWEs, diosyncratically decomposable MWEs; decomposable MWEs - 10 lists of various types (NEs and non-NEs); collocations; and free phrases. The lists are the result of automatic and semi-automatic tagging and classification of the corpus Wiki1,000+ (13.4 million tokens).
    - **Bulgarian Frequency Dictionary** is a lemma frequency dicitionary extracted from the Bulgarian National Corpus (BulNC) that contains 6 domain-specific subcorpora. Thus, 6 domain-specific frequency subdictionaries were developed independently, as well as a general dictionary to combine all domain-specific ones.
- **Technology Tools / Services**
    - **ClauseAlign** –is a tool for alignment of parallel texts at clause level based on a resource light flexible method for clause alignment which combines the Gale-Church algorithm with internally collected textual information.

- **BgMWE Tool for MWE and NE Recognition** is a tool for MWE recognition, categorisation and tagging. It comprises a set of modules developed in Java for corpus processing, annotation, MWE recognition and evaluation. It recognises over 10 different types of MWEs (NEs and non-NEs). The modules can be easily integrated into various systems for corpus processing. BgMWE is generally language independent. It uses text in plain or XML format and also includes a module for format conversion.

- **Web-based Infrastructure for Bulgarian Data Processing** is a highly scalable web service infrastructure that provides easy access to the tools for text processing and annotation of Bulgarian. Three types of access are provided: online access; access via RESTful API; and asynchronous access. The Bulgarian Language Processing Chain includes the following types of text processing and linguistic annotation: sentence segmentation; tokenisation; POS tagging and grammatical annotation; lemmatisation.

- **Bulgarian Word Sense Disambiguation Tool** currently uses 5 independent weak classifiers and an ensemble one that combines all of them. Each of the 6 classifiers provides confidence distribution over the senses for a particular single word or MWE. The current version outperforms the calculated random sense baseline by 24 points.

- **Bulgarian Spell Checker for Mac** – The Bulgarian Spell Checker MacEst for MacOS detects and marks the incorrectly written words in a text and suggests the most probable candidates to correct the errors. MacEst offers a proficiently compiled dictionary, which contains 1.5 mega words.

- **Bulgarian Grammar Checker for Windows** - The Bulgarian Grammar Checker WinEst+ allows users to check and correct texts written in Bulgarian. The formal model for description consists of grammar rules, encoding the (im)possible sequences of grammatical categories, particular lexical items and punctuation. On the basis of a text archive of 450 million words, bi- and tri-grams are generated exemplifying the combinations of grammatical categories that are (im)possible for Bulgarian. These are the basis for the context rules. A simple and effective technology is used for automatic recognition of text positions, where negative sequences of word categories and/or punctuation are expected to be found.

- **Bulgarian Grammar Checker Web Service** – The Bulgarian Spell Grammar WinEst+ is integrated as a web service – both the web service integration and the online grammar checking (as an illustration of the integration) are possible. WinEst+ allows the users to check and correct texts written in Bulgarian on the Internet. The Grammar 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 for correction of texts written in Bulgarian.

resourceName	**resourceShortName**	**resourceLocation**	**resourceType**	**size**	**sizeUnit**	**LingualityType**	**Outside the consortium**
Bulgarian-English clause aligned corpus	BulEnAC	http://dcl.bas.bg/en/corpora_en.html	corpus	30,385	sentence	bilingual	no
Lists of Bulgarian Multiword Expressions	BulMWEs	http://dcl.bas.bg/Resources/MWEs/	LexicalConceptualResource	27,784	multiword units	monolingual	no
Bulgarian Frequency Dictionary	Bulgarian Freq Dictionary	http://dcl.bas.bg/Resources/Frequency/Frequency.zip	exicalConceptualResource	2,142,555	word	monolingual	no
ClauseAlign	ClauseAlign	http://dcl.bas.bg/en/programs_en.html	technologyToolService	-	-	multilingual	no
BgMWE – Tool for MWE and NE Recognition	BgMWE	http://dcl.bas.bg/Tools/MWEs/bg	technologyToolService	-	-	monolingual	no
Web Based Infrastructure for Bulgarian Data Processing	DCLservices	http://dcl.bas.bg/dclservices/registration/	technologyToolService	-	-	monolingual	no
Bulgarian Word Sense Disambiguation Tool	BulWSD	http://dcl.bas.bg/en/programs_en.html	technologyToolService	-	-	monolingual	no
Bulgarian Spell Checker for Mac	MacEst	http://dcl.bas.bg/en/MacEst-en.html	technologyToolService	1.5 mega	word	monolingual	no
Bulgarian Grammar Checker for Windows	WinEst+	http://dcl.bas.bg/est/index_en.php#tabs-5	technologyToolService	-	-	monolingual	no
Bulgarian Grammar Checker Web Service	WebEst+	http://dcl.bas.bg/est/index_en.php#tabs-5	technologyToolService	-	-	monolingual	no

Table 4. Summary of the language resources developed in Bulgaria

3.2. Summary of the language resources developed in Croatia and potentially available to the language engineering community

The basic resources developed in Croatia, many of which are constantly updated, can be classified in the following categories:

- **Monolingual corpora**
  - **Croatian Speech Corpus (CroSpeak Corpus)** is the corpus of recorded Croatian speech covering radio weather forecasts, radio news, read tales, weather dialogs, and TV news (featuring unspontaneous and spontaneous

speech). Overall size of the corpus is 19.35 hours and 227,280 tokens. All utterances have been transcribed following standard Croatian orthography. The corpus has been compiled and processed at the Department of Information Sciences, University of Rijeka.

- **Slovene Web Corpus (slWaC)** is the first version of the Slovene web corpus. It was collected by crawling the whole .si internet domain yielding ca 380 million tokens. The corpus has been lemmatised and MSD-tagged automatically using ToTaLe system by Tomaž Erjavec. The compilation of the corpus is described in the TSD2011 paper *hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene*. The morphosyntactically annotated and lemmatized corpus is distributed under the CC-BY-SA license. 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. The first version is freely accessible for querying at http://faust.ffzg.hr/bonito2/run.cgi/first_form?corpname=slwac.
- **Serbian Web Corpus (srWaC)** is the first version of the Serbian web corpus. It is the work in progress with more than 1 billion of tokens expected. So far 400 million tokens has been collected using the same methodology and tools used for collecting hrWaC and slWaC. The desired size is expected to be reached in 2012-09.

- **X-lingual Croatian corpora**
  - **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 following procedures were applied to resolve existing issues: (1) stricter extraction process – no HTML residues present; (2) language identification on every non-English document – non-English online documents contain English material in case the article was not translated into that language; (3) resolving encoding issues in Croatian and Serbian – diacritics were partially lost due to encoding errors – text was rediacritized. The sentence-aligned language combinations are freely downloadable in TMX or TXT/Moses format. The corpus is published under the CC-BY-SA license.
- **Lexical Conceptual Resources**
  - **Croatian-English Giza++ phrase table** is the Croatian-English translation model built on the basis of several Croatian-English parallel corpora: Croatian-English Parallel Corpus, Croatian-English Parallel Web Corpus (hrenWaC), Croatian-English aligned sentences from SouthEast European Parallel Corpus (SETimes Corpus v2.0).
- **Technology Tools / Services**
  - **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 *hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene*. An implementation (a java

file) is published under the Apache 2.0 licence can be downloaded from http://www.nljubesic.net/upload/ContentExtractor.java. It requires jsoup-1.4.1.jar. A Croatian evaluation sample used in the paper can be downloaded from http://www.nljubesic.net/upload/gold_standard.zip and it is distributed under the CC-BY-SA license.

- **Croatian Academic Spelling Checker (Hascheck)** is one of the oldest Internet services in Croatia. In various forms it acts as a public service and free spelling checker for text written in Croatian language since the spring of 1994. Hascheck’s dictionary database is organized into three sections: (1) Croatian general lexicon, (2) Croatian lexicon of names, (3) English general lexicon. Dictionary database is not static and it is being constantly improved. In the background a working expert system that learns new words from texts submitted for processing. The database is maintained through supervised learning process and currently it exceeds one million types, which all have a been attested in the texts written in Croatian language. The core solutions for spelling checker were written by Šandor Dembitz and the majority of web site processing was written by Gordon Gledec, while the web interface was written by Hrvoje Miholić.

resourceName	**resourceShortName**	**resourceLocation**	**resourceType**	**size**	**sizeUnit**	**LingualityType**	**Outside the consortium**
Croatian Speech Corpus	CroSpeak	http://www.inf.uniri.hr/~ivoi/CROSPEECH/index.htm	corpus	227 280	token	monolingual	yes
Slovene Web Corpus	hrWaC	http://www.nljubesic.net/resources/corpora/hrwac/	corpus	380 000 000	token	monolingual	no
Serbian Web Corpus	srWaC	http://www.nljubesic.net/resources/corpora/srwac/	corpus	400 000 000	token	monolingual	no
SouthEast European Parallel Corpus	SETimes	http://www.nljubesic.net/resources/corpora/setimes/	corpus	43 142 458	token	multilingual	no
Croatian-English Giza++ table	hrenGiza++	http://hnk.ffzg.hr/hrenGiza	lexicalConceptualResource	1 160 274	entry	multilingual	no
Web Content Extractor	WebContentX	http://www.nljubesic.net/resources/tools/webcontentextractor/	technology Tool			multilingual	no
Croatian Academic Spelling Checker	Hascheck	http://hacheck.tel.fer.hr/	technology Tool			multilingual	yes

Table 5. Summary of the language resources developed in Croatia

3.3. Summary of the language resources developed in Hungary and potentially available to the language engineering community

The basic resources developed in Hungary, many of which are constantly updated, can be classified in the following categories:

- **Monolingual (Hungarian) Corpora**
  - **Hungarian National Corpus** is the national corpus of Hungarian language. It is derived into five subcorpora by regional language variants and into five subcorpora by text genres. The subcorpus to be studied can be chosen by any combination of these. That makes the HNC an appropriate tool to study the differences not just between text genres but between language variants. HNC aims to be a representative general corpus of present-day standard Hungarian.
  - **Child Language Corpus** consists of 60 interviews with 4/6-5/6 year-old Hungarian children (from Budapest having different socio-economical backgrounds) with more than 30 hours recording. The interviews include several tasks (picture-based story-telling, telling the rules of a well-known game) and guided conversations. Each interview was conducted by an adult tester. The resource is available in chat (CHILDES) transcription format.
- **Speech Databases**
  - **Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection** contains 2,000 sentences. Each of the 8 speakers read the sentence set. This parallel speech database is used to train HMM based TTS and for unit selection TTS.
  - **Read speech database in Hungarian** - The first automatic TTS based Hungarian weather forecast application (www.metnet.hu) is based on this database containing weather forecast records.
  - **Di-phone database for text-to-speech conversion** - Di-phone database for di-phone based Hungarian TTS. The Profivox Hungarian di-phone TTS uses a database based on this resource.
  - **Hungarian BABEL phonetic and prosodic segmentation and syntactic analysis** is an add-on to Hungarian BABEL speech corpus and includes phoneme level segmentation, prosodic segmentation and annotation for phonological phrases and disambiguated syntactic analysis for 330 sentences/utterances from the Hungarian BABEL.
  - **Hungarian Spontaneous Speech Database (BEA)** is a multi-functional speech database of Hungarian containing various types of spontaneous speech (including conversations), sentence repetitions and reading. This is the largest speech database of Hungarian consisting of about 270 hour recorded speech material by 265 speakers. The number of annotated materials is close to 50%. The recording circumstances of the speech materials are constant, showing a high technical background. All speakers are recorded in the same sound attenuated room at the Research Institute for Linguistics (HAS, Budapest). It allows analysis of the spontaneous speech samples from various acoustic-phonetic and linguistic aspects. In addition, the BEA Database provides a unique possibility for the research on speech technology. The speech

materials of BEA are annotated at various levels of transcription. The current version includes 135 hours of recordings along with the transcribed (in transcriber) and aligned texts by 179 speakers.

- **Tesztel Hungarian Noisy Telephone Speech Corpus** contains noisy speech samples for noise robust ASR in Hungarian, recorded via PSTN and mobile telephone network. The aim was to create a mobile phone voice based Hungarian speech database recorded in noisy environments for testing purposes. The database contains voices of 100 speakers recorded through mobile telephones in noisy environments. The main goal was to test phoneme based recognizers which have been already trained, so the corpus had to be compact, and to cover the specific character of the Hungarian language.
- **Hungarian Child Database for Speech Processing Applications** is useful for speech training and language learning applications for children.
- **BABEL and MRBA sentence modality annotation for Hungarian** – This corpus holds modality annotations for subparts of the Hungarian BABEL and MRBA corpora. It is a useful source for research and analysis or recognition of sentence modality.

- **Lexical Conceptual Resources**
  - **Hungarian Human-Computer Interaction Technologies Multimodal Database** – This is the first and only resource for Hungarian language of aligned text-video-audio segments. The alignment is made by speech units. The audio-visual database recording and annotation project is carried out by the HuComTech (Hungarian Human-Computer Interaction Technologies) multidisciplinary research team involving computer scientists (for digital image processing), computational linguists (for speech processing) and communication experts. The work was carried out between June 2009, and June 2011. The HuComTech (Hungarian Human-Computer Interaction Technologies) project aims at building a multimodal (audio and visual) database of Hungarian language. The research contributes to the knowledge of the interplay of prosodic, verbal and non-verbal features of communicative events by the examination of annotated spontaneous dialogues. Both, formal guided job interviews and informal semi-guided conversations have been recorded with approximately 110 Hungarian university students, resulting in a huge Hungarian audio-visual database complemented with detailed, multi-level multimodal annotation.
  - **ht-online** is a unique lexical database of the most common loanwords in Hungarian language used outside Hungary (collected from 7 regions). The database should be used as a special lexical resource in the Hungarian language tools based on the Hungarian morphology.
  - **Hungarian Concise Dictionary (with sample sentences)** is a unique dictionary of Hungarian language covering 16,000 headwords (entries) followed by frequency data. Each entry describes the most common forms (selected on pragmatical basis) of the headword. The entries are divided into meanings – up to 33,000 carefully selected and stylistically labelled meanings. The dictionary contains sentences brought from real language use and 3,000 phrasems.

- **Technology Tools / Services**
    - **High-speed Unification Morphology (Humor)** is a reversible, string-based unification approach for lemmatizing and disambiguating language data that has been used for both, language corpus analysis and creation of a variety of linguistic software applications such as spell-checkers. The system is language-independent, allowing multilingual applications for a variety of language types. Its Hungarian version, the largest and most precise implementation, contains nearly 100,000 stems. The system has been tested rigorously by both linguists and end-users of word-processing tools. Humor-based linguistic modules have been licensed by major software producers, and the lemmatizer has been used in lexicographic research since 1991. One tool provides disambiguation, tagging, and parsing functions. The system can describe various natural languages, including both Eastern European and non-Eastern European languages. Several Humor subsystems for different purposes (lemmatizing, hyphenating, spell-checking/correcting, grammar checking) are commercially available, and have been built into several major word-processing and full-text retrieval systems. An inflectional thesaurus and a series of intelligent bilingual dictionaries have also been developed.
    - **Graphical query interface for Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection** is designed to allow fast access, search and statistical query possibilities and functionality for the Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection, thus it acknowledges advanced phonetic/speech technology research on the corpus.
    - **Multilingual speech segmentation tool** is used for phoneme segmentation of utterances for 6 languages: English, French, Italian, Spanish and Hungarian.
    - **Sentence modality recognizer** (based on Hungarian and German speech models) can be used in speech recognition and understanding.

resourceName	**resourceShortName**	**resourceLocation**	**resourceType**	**size**	**sizeUnit**	**Linguality Type**	**Outside the consortium**
Hungarian National Corpus	HNC	corpus.nytud.hu/hnc	corpus	187,600,000	token	monolingual	no
Hungarian Spontaneous Speech Database	BEA	HASRIL	corpus	270	hour	monolingual	no
Child language corpus	CHILC	HASRIL	corpus	60	interview	monolingual	no
Hungarian Human-Computer Interaction Technologies Multimodal Database	HUCOMTECH	University of Debrecen	multimodal corpus	50	hour	monolingual	yes

Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection	ParallelSpeech-hu	BME-TMIT	corpus	25	hour	monolingual	no
Read Speech Database for Hungarian	ReadSpeech-hu	BME-TMIT	corpus	10	hour	monolingual	no
Hungarian BABEL phonetic and prosodic segmentation and syntactic analysis	BABEL-Addon1	BME-TMIT	corpus	330	utterance	monolingual	no
Di-phone database for text-to-speech conversion in Hungarian	Di-phone-hu	BME-TMIT	corpus	1,646	second	monolingual	no
Tesztel Hungarian Noisy Telephone Speech Corpus		BME-TMIT	corpus	100	speaker	monolingual	no
A Hungarian Child Database for Speech Processing Applications		BME-TMIT	corpus	72	speaker	monolingual	no
BABEL and MRBA sentence modality annotation for Hungarian		BME-TMIT	corpus	50	speaker	monolingual	no
ht-online	ht-online	Termini Research Network	lexicalConceptualResource	4,000	entry	monolingual	yes
Hungarian Concise Dictionary (with sample sentences)	HCD	TINTA Publishing House	lexicalConceptualResource	16,000	entry	monolingual	yes
High-speed Unification Morphology	HUMor	MorphoLogic Ltd.	technologyToolService	100,000	entry	monolingual	yes
Graphical Query Interface for Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection	-	University of Debrecen	technologyToolService	-	other	monolingual	yes
Multilingual Speech Segmentation Tool	-	BME-TMIT	technologyToolService	-	oher	multilingual	no
Sentence Modality Recognizer	-	BME-TMIT	technologyToolService	-	other	bilingual	no

Table 6. Summary of the language resources developed in Hungary

3.4. Summary of the language resources developed in Poland and potentially available to the language engineering community

The basic resources developed in Poland, many of which are constantly updated, can be classified in the following categories:

- **Monolingual Corpora**
    - **Składnica** is the result of the Polish Ministry of Science and Higher Education research grant (ended in October 2011) on construction of a treebank for Polish by using automatic syntactic analysis. The resource is a treebank of Polish constituents created automatically and then manually corrected.
    - **The Corpus of Polish Summaries** aims at collecting human-written summaries of 154 texts, each text sized between 1,000 and 4,000 words and extracted from Rzeczpospolita Corpus (http://www.cs.put.poznan.pl/dweiss/rzeczpospolita) – a corpus of press articles from the website of the *Rzeczpospolita* newspaper. The set of articles contains articles published since year 1993 to 2002 and is yet not freely available. A set of frequently represented text categories in the Rzeczpospolita Corpus was chosen: economics, law, news from Poland, culture, sport, science and technology, opinions. The corpus contains two types of summaries: abstractive and extractive. Each text is going to have 3 summaries of both types, varying in length: a 20%, 10% and 5% summary (in terms of word count of original text). Abstractive summaries are written by the annotators as a free text, extractive summaries are created by selecting unconstrained fragments of the original text (following guidelines) in terms of single character as the smallest possible selection. The 10% extractive summary contains only a subset of selections in 20% extractive summary, etc.
    - **Learner Speech Database (PELSC)** contains samples of spoken learner English from the PELCRA Learner English Corpus. The database contains transcriptions of Poles speaking in English and Polish on a variety of informal topics. The transcriptions are time-aligned at the level of utterances with the underlying recordings, most of which are studio-quality and uncompressed. Possible NlpApplications include the improvement of speech recognition systems made for speakers of English with a Polish accent.
    - **SNUV voice recognition speech database (SNUV)** is a spelling and number and recognition speech database composed of 200 hours of recordings of Polish speakers reading numbers and spelling words, recorded in 22050 kHz, 16-bit *.wav files. It includes a transcription of the recordings in text format, encoded in the UTF-8 standard. The purpose of the resource is to enable the creation of automatic speech recognition (ASR) tools to allow the user to spell out a word or a number to be recognized. SNUV is potentially the largest available Polish speech recognition database, which can be released under a CC-license.
    - **PELCRA Time-Aligned Spoken Corpus** is the largest collection of transcriptions of naturally occurring conversational Polish that has been compiled by the PELCRA team at the University of Łódź. It contains over 40

hours of conversation recorded in an informal setting. The transcriptions have been time-aligned with the original recordings at the level of utterances.

  - **Paralela DB** is a multilingual parallel corpus containing texts of CORDIS news database, RAPID press release of the EU, press releases of the European Parliament and of the European Southern Observatory. Except for Polish which is obligatory, the database covers more than 20 other languages. The process of converting, processing and exporting parallel resources encoded in a variety of formats is facilitated by the use of a central relational database system (named *Paralela*) to which text collections are imported. The Paralela database is used to store bibliographic, structural and alignment information, and is designed to handle multiple alignments of the same collection. Once the variously encoded collections are converted and normalised, they can be processed and exported into more uniform and standard formats used for the exchange of parallel corpora and translation memories.

- **Bilingual Corpora**
  - **Parallel English-Polish Corpus** collected within the ATLAS (Applied Technology for Language-Aided CMS) project comprises contemporary works, both literary and from restricted language domains, from accounting, computing, politics, biology or physics, music, sociology or fine arts. The corpus is manually aligned by trained annotators on the sentence level, with a custom methodology developed to represent all non-trivial translation equivalence types (deletions, insertions, splits, merges, paraphrases, etc.). Serving as a basis for the provision of language models and other deliverables of the project, the corpus is exported in formats following industry standards and best practices, as TEI P5-compliant XML files with custom extensions to mark complex translation equivalence types, as well as in the XLIFF and TMX formats.
  - **Redistributable Polish-Russian Corpus** is currently being created to reach 50 million words, 50% of Polish originals translated into Russian and 50% vice versa. The core of the resources consists of the literary classics of the nineteenth century and contemporary works which are the most popular in the neighbouring country. The corpus contains press texts and their translations, as well as legal texts. The texts are annotated according to the DTDs of the National Corpus of Polish and the Russian National Corpus. A morphosyntactic search is possible, although the standards of the two national corpora differ in a number of grammatical classes and categories.

- **Lexical Conceptual Resources**
  - **LFG Grammar of Polish** is currently being constructed by making extensive reuse of existing language resources for Polish. Its constituent structure (c-structure) is based on a DCG grammar of Polish and the functional structure (f-structure) was mainly inspired by the available HPSG analyses of Polish. Valence information from the dictionary which accompanies the DCG grammar was converted so that subcategorisation is stated in terms of grammatical functions rather than categories; additionally, missing valence frames may be extracted from the treebank. The obtained grammar is evaluated using constructed test suites (half provided by previous grammars) and the treebank.

- **Formal Grammar of Polish (GFJP)** is the most extensive and most detailed formal grammar of Polish expressed as a metamorphosis grammar with several extensions, e.g., allowing for permuting phrases. Syntactic units are represented by terms of parameters formalizing various grammatical features of those units. Rules of the grammar define particular units as sequences of other units and establish correspondences between grammatical features (unification). Agreements are accounted for by parameter matching used an extensive set of parameters. The values a given unit is assigned, be it from the top ("syntactic" features) or from the bottom ("lexical" features), to spread down the syntactic tree, reaching most of its constituents. Rules defining different syntactic units (sentences or phrases) follow one format. The grammar has the ambition to define the whole language to cover most structures of Polish.

- **Syntatic-generative dictionary of Polish verbs** has been published on paper in the 1980s and 1990s. Then, its computer implementation in the form of a MS Access database was created. Currently, after IPR clarification, a better representation format is being constructed for the resource.

- **Polish OpenCYC lexicon** covers translation of a substantial portion of the conceptual part of the OpeCyc ontology into Polish. The work is concentrated on precise identification of the Polish lexemes that correspond to the English concepts. Special attention is paid to multiword entries. The part of Cyc that was selected for translation is roughly equivalent to the contents of UMBEL – an ontology extracted from Cyc specifically for various NLP tasks. The result of the translation will cover mappings to Polish inflectional dictionary. This lexicon will be integrated back into OpenCyc, after the work is finished.

- **Polish-English Wikipedia NE dictionaries** are Wikipedia-derived English-Polish and Polish-English thematic dictionaries that can be used in NLP applications, e.g., for tagging media-related texts with information about their content. The dictionaries are based on existing Wikipedia categories, but they have been manually checked for inappropriately-placed entries. Subjects that are covered includeUS universities, world cities and villages, Polish artists, journalistsm, scientists, companies, organisations, etc. The dictionaries are stored in the RDF (Resource Description Framework) program, which is a method for conceptual description or modeling of information that allows storage of additional information. The categories do not reflect the exact Wikipedia structure, but conceptual relations.

- **Technology Tools / Services**
  - **Lexeme Forge** is a Web-based tool used to manage creation of morphological dictionaries for inflectional languages. The system manages a database of lexemes and allows editing of their descriptions to define their inflectional paradigms. The database is modelled after the Grammatical Dictionary of Polish, in particular using its inflectional patterns. The system allows attachment of various labels to the lexemes. Besides typical dictionary labels such as informal or dated, special labels are used for excluding some forms from spell-checking dictionaries. Thus, a special variant of the dictionary can be generated without containing some theoretically correct but extremely infrequent words (i.e., potential false negatives in spell-checking). Moreover, the system makes it possible to specify a classification scheme (or several

classification schemes), which the lexemes are to follow. This mechanism is currently used to classify lexemes into common and proper names (with some subclasses).

- **Slowal** is a Web tool designed for creating valence dictionaries based on the format presented by Filip Skwarski. It describes lemmas by a list of individual frames presented as tables which can be expanded by adding to them new positions, arguments, series of characteristics and examples showing their usage. The tool provides user group management (Guests – add notes to created lemmas, Lexicographers – responsible for expanding existing lemmas, Superlexicographers – responsible for checking correctness of lexicographers work, managing vocabularies and adding new lemmata). Slowal implements a list of features helping in creating and expanding lemmas, e.g., looking for similar lemmas, validation of created frames, series of filters to help find lemmas using specific position or arguments and much more. Such created vocabularies can be imported from text format. Slowal is implemented using Django framework.

- **Lakon** is a Polish extractive summarizer using algorithms based on salient sentence selection, namely: heuristic evaluation of position of sentences in paragraphs, word weighting schema tf-idf and okapi bm25 as well as lexical chains combined with thesaurus use. The quality of the automatically generated summaries have been evaluated against a corpus of manually created summaries of selected press articles.

- **Świgra** is a Prolog parser implementing Świdziński's Formal Grammar of Polish. Świgra uses a bottom-up parsing strategy, which for Polish proved to be superior to the top-down strategy. The parser builds a shared parse forest, which is not only the result, but also succeeds in avoiding unnecessary recomputation. The rules of the grammar are not interpreted at the runtime but are compiled to Prolog clauses.

- **Anotatornia** is a tool for manual on-line annotation of corpora at various linguistic levels. The levels currently implemented are: word-level and sentence-level segmentation, morphosyntax, word sense disambiguation. Anotatornia implements sophisticated mechanisms of the management of texts, annotators and conflicts.

- **Ruler** is a rule-based coreference resolver for Polish. The implemented module uses standard best-first entity-based model based on syntactic constraints (elimination of nested nominal groups), syntactic filters (elimination of syntactic incompatible heads), semantic filters (wordnet-derived compatibility) and selection (weighted scoring). Syntactic properties are obtained from Spejd and its morphological component Morfeusz SGJP which produce NP chunks with detailed morphosyntatic information. Semantic properties are currently based on plWordNet.

- **PolSumm** is a Polish document summarizer combining elements of a linguistic transformation of the text with statistical methods and information retrieval.

- **VOICE LAB Automated Speech Recognition (ASR) engine** enables recognition of natural speech. The ASR supports an industry standard known as Speech Recognition Grammar Specification (SRGS). The engine has been optimized for use in navigation of information kiosks, mobile applications, switch-boards or call centres supporting human operators, as well as in voice search. The ASR can be used as a service or as a standalone, on-site installation. The acoustic models have been optimized for Polish. With appropriate training, it can be used for any language as the core technology is language independent. The engine works on every Linux distribution, preferably a 64 bit one.

- **The Language Detector (LDetect)** is a Java tool for detecting the language of an arbitrary stretch of text developed by the PELCRA team at the University of Łódź and available under the GPL licence. The first version 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.

resourceName	**resourceShortName**	**resourceLocation**	**resourceType**	**size**	**sizeUnit**	**LingualityType**	**Outside the consortium**
Składnica	Składnica	-	corpus	8,227	sentence	monolingual	yes/no
The Corpus of Polish Summaries	SummaryCorpus	-	corpus	yet unknown	text	monolingual	yes/no
The Parallel English-Polish Corpus	ParallelCorpus	-	corpus	3,000,000	word each side	bilingual	yes/no
Redistributable Polish-Russian Corpus	DistrPLRU	-	corpus	yet unknow	word	bilingual	yes/no
Learner Speech Database	PESLC	-	corpus	50,000	word	monolingual	no
SNUV voice recognition speech database	SNUV	http://snuv.pl	corpus	200	hour	monolingual	no
PELCRA Time-Aligned Spoken Corpus	TASC	http://pelcra.pl/tasc	corpus	40	hour	monolingual	no
Paralela DB	Paralela	http://pelcra.pl/paralela	corpus	50,000,000	word	multilingual	yes/no
LFG Grammar of Polish	LFGGrammarPL	-	lexicalConceptualResource	yet unknown	entry	monolingual	no
Formal Grammar of Polish	GFJP	-	lexicalConceptualResource	460	rule	monolingual	yes/no

Syntatic-Generative Dictionary of Polish Verbs	SSGCP	-	lexicalConceptualResource	10,559	verb entry	monolingual	yes
Polish OpenCYC lexicon	OpenCYCPL	-	lexicalConceptualResource	yet unknown	word	bilingual	yes/no
Polish-English Wikipedia NE dictionaries	NERDict	http://pelcra.pl/res/ecl-dictionaries	lexicalConceptualResource	-	entry	bilingual	no
Lexeme Forge	LexemeForge	-	technologyToolService	-	-	monolingual	no
Slowal	Slowal	-	technologyToolService	-	-	monolingual	no
Lakon	Lakon	-	technologyToolService	-	-	monolingual	no
Świgra	Świgra	-	technologyToolService	-	-	monolingual	yes/no
Anotatornia	Anotatornia	-	technologyToolService	-	-	monolingual	yes/no
Ruler	Ruler	-	technologyToolService	-	-	monolingual	yes/no
PolSumm	PolSumm	-	technologyToolService	-	-	monolingual	yes
VOICE LAB Automated Speech Recognition (ASR) engine	VLASR	http://www.voicelab.pl/	technologyToolService	-	-	monolingual	yes
Language Detector	LDetect	http://pelcra.pl	technologyToolService	-	-	multilingual	no

Table 7. Summary of the language resources developed in Poland

3.5. Summary of the language resources developed in Serbia and potentially available to the language engineering community

The basic resources developed in Serbia, many of which are constantly updated, can be classified in the following categories:

- **Multimedia Corpora**
  - **Media Multimedia Archive Ebart** is a video archive that contains several hundred thousand broadcasts from the most important central TV stations and some local TV stations published since 2005. They are grouped on various criteria (thematic, persons, etc.). A large number of them are transcribed to text. Media Archive Ebart was developed by the Ebart Archive, Belgrade. The EbartArchive full-text database contains articles from 27 daily and weekly newspapers, as well as articles from 16 special newspaper supplements and 17 local newspapers published throughout Serbia. Topics covered include Serbian current events, politics, economics, science, culture, and public life. With archives from 2003 to the present, the database contains approximately 4 million fully indexed articles.

- **Monolingual Text Corpora**
  - **Named Entities Evaluation Corpus for Serbian (SrpNE-evaluation)** consists of approx. 3,000 short news in which named entities were automatically tagged and manually checked. Named entities tagged are: persons, person roles and functions, temporal expressions, mount expressions (including measures and money expressions) and organizations.
  - **Semantically Tagged Corpus of Contemporary Serbian (preliminary version)** was semantically tagged on the basis of some semantic attributes associated to lemmas in Serbian e-dictionaries, as well as on Serbian Wordnet.

- **Bilingual and Mulilingual Text Corpora (with Serbian as one language)**
  - **Serbian-English Aligned Literary Corpus** consists of Serbian literary texts translated to English.
- **Lexical Conceptual Resources**
  - **Terminological Database for Geology (GeolISSTerm)** is an electronic dictionary of geologic terms based on a special-purpose taxonomy of basic geologic concepts and terms. GeolISSTerm is an elementary electronic resource in the process of domain formation in the Geologic Information System of Serbia (GeolISS). It is the core of GeolISS through which validation, classification and specification of attributes of the observed and the interpreted takes place.

- **Technology Tools / Services:**
  - **Emotion Classification of Serbian Texts** is a system based on ontology built specially to function as an emotion classifier. The application is realized on Csharp Net Framework platform. It can be tested on texts in .html and .txt

formats and accepts both Cyrillic and Latin scripts. Text files can be manually pasted, uploaded from a local system or used directly from a given URL address on Web.

- **Named entities module for Serbian (SrpNE-module)** is module for named entity recognition and tagging based on Serbian morphological e-dictionaries and a large collection of Finite-State Transducers (in the form of cascades). It recognizes and tags: persons, person roles and functions, temporal expressions, mount expressions (including measures and money expressions) and organizations. The module is integrated in a web service and tags NEs in texts uploaded by users.
- **A web tool for aligned text search** is used for effective search of aligned and annotated texts. It is especially designed for texts in which named entities were tagged. Its purpose is to compare annotation of NEs in aligned text and for that purpose a language independent classification schema for NEs is used.
- **Web applications (NE extraction from web pages)** are a set of web tools for extraction of proper names from categories given in Wikipedia for English, French, Serbian, Polish.

- **Language description:**
  - **Language Model for Serbian** is produced on the basis of the large newspaper corpus (approx. 4 million articles) using the standard methodology for such models.

resourceName	resourceShortName	resourceLocation	resourceType	size	sizeUnit	LingualityType	Outside the consortium
Multimedia Ebart Archive	Ebart Archive	http://www.arhiv.rs/	speech corpus	-	article	monolingual	yes
Named Entities Evaluation Corpus for Serbian	SrpNE-evaluation	http://korpus.matf.bg.ac.rs	corpus	150,000	word	monolingual	no
Semantically tagged Corpus of Contemporary Serbian (preliminary version)		http://korpus.matf.bg.ac.rs	corpus	-	word	monolingual	no
Serbian-English Aligned Literary Corpus	-	http://www.ff.uns.ac.rs/	corpus	-	word	bilingual	yes
Terminological Database for Geology	GeolISS Term	http://www.rgf.bg.ac.rs/	lexicalConceptualResource	3,500	concept	bilingual	yes
Emotion classification of Serbian Texts	-	http://korpus.matf.bg.ac.rs	technologyToolService	-	-	monolingual	no
Named entities module for Serbian	SrpNE-module	http://korpus.matf.bg.ac.rs	technologyToolService	-	-	monolingual	no
A web tool for aligned text seach	-	http://korpus.matf.bg.ac.rs	technologyToolService	-	-	multilingual	no

Web applications (NE extraction from web pages)	-	http://korpus.matf.bg.ac.rs	technologyToolService	-	-	multilingual	no
Language Model for Serbian	-	-	language description	-	-	monolingual	yes

Table 8. Summary of the language resources developed in Serbia

3.6. Summary of the language resources developed in Slovakia and potentially available to the language engineering community

The basic resources developed in Slovakia, many of which are constantly updated, can be classified in the following categories:

- **Lexical Conceptual Resources**
  - **Database of Root Morphemes** provides alternative approach to morphology analysis. It contains 67,000 linguistic units with deep morphematic linguistic analysis. It has been compiled at the Prešov University in Prešov and has been used as a basis for a published Slovník koreňových morfém slovenčiny (M. Sokolová et al.). ISBN 9788080683191.
  - **Dictionary of Slovak Adjective Collocations** provides an overview of the combinatorial behaviour of words and contains collocation profiles of the most frequent Slovak adjectives. The combinatorial potentials of word forms of a word are the basis for the creation of so-called collocational templates which the patterns of collocations are based on. The dictionary is currently being compiled (presently, it contains collocation profiles of 140 adjectives). The dictionary is being created at the University of St. Cyril and Methodius in Trnava, with input from the Ľ. Štúr Institute of Linguistics.
  - **Dictionary of German-Slovak Collocations** provides confrontational overview of the combinatorial behaviour of words in bilingual comparison. The database consists of German collocations (currently 440 profiles) with Slovak equivalents The dictionary is being created at the University of St. Cyril and Methodius in Trnava.
  - **Multimodal Multilingual Dictionary of Gestures (DiGest)** contains a database of extra-verbal expressions. Its current version contains several hundreds of gestures represented by a still image, a description of the gesture and its meaning, and optional sound and video records. The current version includes language and culture dependent content for American English, Slovak, Italian, and Mongolian. Entries for Japanese, Chinese, and Hungarian are also included. The database has been compiled at the Institute of Informatics, Slovak Academy of Sciences.
- **Technology Tools / Services:**
  - **Language model prim-5.0-inf** is a language model from the Slovak National Corpus. The model is in iARPA format, using written-bell smoothing. It was

created by the IRSTLM Tooklit. The model is lower-cased and has been released with the contribution of the EuroMatrixPlus project.

- **Language model prim-5.0-vyv** is a language model of balanced language built on the balanced Slovak corpus. The model is in iARPA format, using witten-bell smoothing. It was created by the IRSTLM Tooklit. The model is lower-cased and has been released with the contribution of the EuroMatrixPlus project.
- **Language model prim-5.0-sane** is a language model from the Slovak National Corpus. The model is in iARPA format, using written-bell smoothing. It was created by the IRSTLM Tooklit. The model is lowercased. It has been released with the contribution of the EuroMatrixPlus project.

resourceName	**resourceShortName**	**resource Location**	**resourceType**	**size**	**sizeUnit**	**LingualityType**	**Outside the consortium**
Database of Root Morphemes	Database of root morphemes	Prešov University	lexicalConceptualResource	67,000	root morpheme	monolingual	yes
Dictionary of Slovak Adjective Collocations	Dictionary of Slovak Adjective Collocations	University of St. Cyril and Methodius, Trnava	lexicalConceptualResource	140	entry	monolingual	yes
Dictionary of German-Slovak Collocations	Dictionary of German-Slovak Collocations	University of St. Cyril and Methodius, Trnava	lexicalConceptualResource	440	entry	bilingual	yes
Multimodal Multilingual Dictionary of Gestures	DiGest	Institute of Informatics, Slovak Academy of Sciences	lexicalConceptualResource	324	entry	multilingual	yes
Language model prim-5.0-inf	Language model prim-5.0-inf	LSIL	technology ToolService	515,000,000	token	monolingual	yes
Language model prim-5.0-vyv	Language model prim-5.0-vyv	LSIL	technology ToolService	247,000,000	token	monolingual	yes
Language model prim-5.0-sane	Language model prim-5.0-sane	LSIL	technology ToolService	733,000,000	token	monolingual	yes

Table 9. Summary of the language resources developed in Slovakia

4. Conclusions

During the reported period, 73 resources were developed, updated, or contacted. A third of the resources are corpora (14 text and 13 audio or multimedia). 16 are lexical/conceptual databases, while technology tools / services are 30 (or more than 40%). Most of the resources (53 out of 73) are monolingual (distributed among the different languages), while 19 are bilingual or multilingual.
Finally, thirty from the resources are identified outside the consortium (45%).

Resources per Coutry	**Total**	**By Resource type**				**By Linguality**			
		Text Corpora	**Audio Corpora**	**Lexical / Conceptual Database**	**technology tool / service**	**Monolingual**	**Bilingual**	**Multilngual**	**Outside the consortium**
Bulgaria	10	1	-	2	7	8	1	1	-
Croatia	7	3	1	1	2	3	-	4	2
Hungary	17	2	8	3	4	15	1	1	5
Poland	22	5	3	5	9	16	4	2	13
Serbia	10	3	1	1	5	6	2	2	4
Slovakia	7	-	-	4	3	5	1	1	7
Total	**73**	**14**	**13**	**16**	**30**	**53**	**9**	**10**	**31**

Table 10. Summary of the reported language resources

5. Annex

5.1. Bulgarian language resources detailed specification

resourceName	Bulgarian-English clause aligned corpus
resourceShortName	BulEnAC
downloadLocation	
dateCreation	2012
projectPartner	Institute for Bulgarian Language
iprHolder.organizationName	Institute for Bulgarian Language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	underNegotiation
license	
resourceLocation	http://dcl.bas.bg/en/corpora_en.html
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	nlpApplications
actualUse	nlpApplications
description	The corpus consists of 363,402 tokens altogether (174,790 for Bulgarian and 188,612 for English) distributed over five thematic domains: fiction (21.4%), news (37.1%), administrative (20.5%), science (11.2%) and subtitles (9.8%). The purpose of using a general testing corpus with texts from a variety of domains is to investigate method performance in a wider range of contexts. Both Bulgarian and English parts of the corpus are first automatically segmented and then aligned at sentence level. Bulgarian sentences are manually or semi automatically split into clauses and for the English texts a pre-trained OpenNLP parser is used to determine clause boundaries followed by manual expert verification and post-editing (the task of automatic clause splitting falls outside the scope of the present study). Subsequently, manual clause alignment is performed.
relevantPublications	S. Koeva, B. Rizov, E. Tarpomanova, T. Dimitrova, R. Dekova, I. Stoyanova, S. Leseva, H. Kukova, and A. Genov. Application of clause alignment for statistical machine translation. In Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6), 12 July 2012, Jeju, Korea, 2012.
resourceType	corpus
mediaType	text
lingualityType	bilingual
languageId	bg
Size	30,385
sizeUnit	sentence

resourceName	Lists of Bulgarian Multiword Expressions
resourceShortName	BulMWEs
downloadLocation	http://dcl.bas.bg/Resources/MWEs/lists.zip
dateCreation	2012
projectPartner	Institute for Bulgarian Language
iprHolder.organizationName	Institute for Bulgarian Language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	available-restricted use
license	
resourceLocation	http://dcl.bas.bg/en/dictionaries_en.html
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	nlpApplications
actualUse	nlpApplications
description	The set of 13 lists comprises of 27,784 entries (MWEs and phrases) divided into the following categories: Non-decomposable MWEs. Idiosyncratically decomposable MWEs. Decomposable MWEs - 10 lists of various types (NEs and non-NEs) where classification is based on the degree of idiomaticity and compositionality. Collocations which are not lexical units and thus are not classified as MWEs. Free phrases. The lists of Multiword expressions are result of automatic and semi-automatic tagging and classification of the corpus Wiki1000+ (13.4 million tokens).
relevantPublications	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)
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	monolingual
languageId	bg
size	27,784
sizeUnit	multiWordUnits

resourceName	Bulgarian Frequency Dictionary
resourceShortName	Bulgarian Freq Dictionary
downloadLocation	http://dcl.bas.bg/Resources/Frequency/Frequency.zip
dateCreation	2012
projectPartner	Institute for Bulgarian Language
iprHolder.organizationName	Institute for Bulgarian Language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla

contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	available-restricted use
license	
resourceLocation	http://dcl.bas.bg/en/dictionaries_en.html
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	humanUse nlpApplications
actualUse	humanUse nlpApplications
description	Bulgarian Frequency Dictionaries are lemma frequency dicitionaries 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.
relevantPublications	
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	monolingual
languageId	bg
size	2,142,555
sizeUnit	word

resourceName	ClauseAlign – tool for alignment of parallel texts at clause level
resourceShortName	ClauseAlign
downloadLocation	
dateCreation	2012
projectPartner	Institute for Bulgarian language
iprHolder.organizationName	Institute for Bulgarian language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	underNegotiation
license	
resourceLocation	http://dcl.bas.bg/en/programs_en.html
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	nlpApplications
actualUse	nlpApplications
description	The ClauseAlign is a resource light flexible method for clause alignment which combines the Gale-Church algorithm with internally collected textual information. The method does not resort to any pre-developed linguistic resources which makes it very appropriate or resource light clause alignment. A combination of the method with the original Gale-Church algorithm (1993) is applied for clause alignment.

relevantPublications	S. Koeva, B. Rizov, E. Tarpomanova, T. Dimitrova, R. Dekova, I. Stoyanova, S. Leseva, H. Kukova, and A. Genov. Application of clause alignment for statistical machine translation. In Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6), 12 July 2012, Jeju, Korea, 2012.
resourceType	technologyToolService
mediaType	text
lingualityType	multilingual
languageId	
size	
sizeUnit	

resourceName	BgMWE – tool for MWE and NE recognition
resourceShortName	BgMWE
downloadLocation	http://dcl.bas.bg/Tools/MWEs/bgMWE.jar
dateCreation	2012
projectPartner	Institute for Bulgarian language
iprHolder.organizationName	Institute for Bulgarian language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	available-restricted use
license	GPLv3
resourceLocation	http://dcl.bas.bg/en/programs_en.html
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	nlpApplications
actualUse	nlpApplications
description	BgMWE is a tool for MWE recognition, categorisation and tagging. It comprises a set of modules developed in Java for corpus processing, annotation, MWE recognition and evaluation. It recognises over 10 different types of MWEs (NEs and non-NEs) with respect to their degree of idiomaticity and compositionality. The modules can be easily integrated into various systems for corpus processing. BgMWE is generally language independent although it is tested only for Bulgarian. It uses text in plain or XML format and also includes a module for format conversion.
relevantPublications	Stoyanova, Ivelina. PhD thesis: Automatic recognition and annotation of compound lexical units in Bulgarian (in Bulgarian). Lists of MWE of different categories.
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	
size	
sizeUnit	

resourceName	Web based infrastructure for Bulgarian data processing
resourceShortName	DCLservices
downloadLocation	http://dcl.bas.bg/dclservices/registration/
dateCreation	2012
projectPartner	Institute for Bulgarian Language
iprHolder.organizationName	Institute for Bulgarian Language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	available-restricted use
license	Other
resourceLocation	http://dcl.bas.bg/dclservices/registration/
distributionAccessMedium	webExecutable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Director
foreseenUse	nlpApplications
actualUse	nlpApplications
description	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. RESTful API access is suitable for software developers who can integrate the processing tools in high level applications. Asynchronous access is aimed at 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 Bulgarian Language Processing Chain includes the following types of text processing and linguistic annotation: sentence segmentation; tokenisation; POS tagging and grammatical annotation; lemmatisation.
relevantPublications	
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	
size	
sizeUnit	

resourceName	Bulgarian word sense disambiguation tool
resourceShortName	BulWSD
downloadLocation	
dateCreation	2012
projectPartner	Institute for Bulgarian language
iprHolder.organizationName	Institute for Bulgarian language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	underNegotiation
license	

resourceLocation	http://dcl.bas.bg/en/programs_en.html
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	nlpApplications
actualUse	nlpApplications
description	The Bulgarian word sense disambiguation tool currently uses 5 independent "weak" classifiers and an ensemble one that combines all of them. Each of the 6 classifiers provides confidence distribution over the senses for a particular single word or MWE (list of pairs: <sense, confidence>, where the sum of the confidences is 1, are generated). Two of the classifiers - a Lesk and a Degree implementation – are knowledge-based. These algorithms disambiguate words using information encoded in BulNet and the context of the word in the corpus. Two other disambiguators are Hidden Markov Model-based – one for forward and one for backward processing of the sequences in the text. The fifth weak classifier, a frequency-based one, assesses the confidence for a particular sense according to its frequency in BulSemCor. The ensemble classifier uses a weighted sum of the five weak ones. The current version outperforms the calculated random sense baseline by 24 points. The ensemble disambiguator shows a good overall improvement in terms of precision outperforming the best of the weak classifiers by approximately 5 points (~65% vs. ~60%). Although some of the algorithms process part of the words in a given text, the coverage of the system is almost 100%, and precision is ~65% (vs ~40% for random sense).
relevantPublications	
resourceType	technologyToolService
mediaType	text
linguality Type	monolingual
languageId	bg
size	
sizeUnit	

resourceName	Bulgarian Spell Checker for Mac
resourceShortName	MacEst
downloadLocation	http://dcl.bas.bg/sites/default/files/webfm/MacEst/MacEst1-1.0-beta1.dmg
dateCreation	2010
projectPartner	Institute for Bulgarian Language
iprHolder.organizationName	Institute for Bulgarian Language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	available-restrictedUse
license	
resourceLocation	http://dcl.bas.bg/en/MacEst-en.html
distributionAccessMedium	downloadable
restrictionsOfUse	noModifications
licenseSignatory.Person.position	director

foreseenUse	human use
actualUse	human use
description	The Bulgarian spell checker MacEst for MacOS detects and marks the incorrectly written words in a text and suggests the most probable candidates to correct the errors. MacEst offers the entire potential of 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.
relevantPublications	
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	bg
size	1.5 mega
sizeUnit	word

resourceName	Bulgarian Grammar Checker for Windows
resourceShortName	WinEst+
downloadLocation	
dateCreation	2012
projectPartner	Institute for Bulgarian language
iprHolder.organizationName	Institute for Bulgarian language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	available-restrictedUse
license	
resourceLocation	http://dcl.bas.bg/est/index_en.php#tabs-5
distributionAccessMedium	downloadable
restrictionsOfUse	noModifications
licenseSignatory.Person.position	director
foreseenUse	human use
actualUse	human use
description	The Bulgarian Grammar Checker WinEst+ allows users to check and correct Bulgarian texts. The formal model consists of grammar rules, encoding the (im)possible sequences of grammatical categories, particular lexical items and punctuation. On the basis of a text archive of 450 million words, bi- and tri-grams are generated exemplifying the combinations of grammatical categories that are (im)possible for Bulgarian and are the basis for the contexts rules. A simple but effective technology is used for automatic recognition of text positions, where negative sequences of word categories and/or punctuation are to be found.
relevantPublications	
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	bg
size	
sizeUnit	

resourceName	Bulgarian Grammar Checker Web Service
resourceShortName	WebEst+
downloadLocation	
dateCreation	2012
projectPartner	Institute for Bulgarian Language
iprHolder.organizationName	Institute for Bulgarian Language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	available-restrictedUse
license	
resourceLocation	http://dcl.bas.bg/est/index_en.php#tabs-5
distributionAccessMedium	webExecutable
restrictionsOfUse	noModifications
licenseSignatory.Person.position	Director
foreseenUse	human use
actualUse	human use
description	The Bulgarian Spell Grammar WinEst+ is integrated as a web service – both the web service integration and the online grammar checking (as an illustration of the integration) are possible. WinEst+ allows users to check and correct Bulgarian texts on the Internet. The Grammar Checker web service can be used in different blogs, chat forums, online shops, media, and everywhere for creation of Internet content, so it will assist the correct writing of Bulgarian texts.
relevantPublications	
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	bg
size	
sizeUnit	

5.2. Croatian language resources detailed specification

resourceName	Croatian Speech Corpus
resourceShortName	CroSpeak Corpus
downloadLocation	http://www.inf.uniri.hr/~ivoi/CROSPEECH/index.htm
dateCreation	2011
projectPartner	FFZG
iprHolder.organizationName	University of Rijeka
contact.Person.surname	Ipšić
contact.Person.givenName	Ivo
contact.Person.email	ivoi@inf.uniri.hr
DistributionInfo	available-restricted
license	
resourceLocation	http://www.inf.uniri.hr/~ivoi/CROSPEECH/index.htm
distributionAccessMedium	not yet available for internet access
restrictionsOfUse	following the license
licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
description	Croatian Speech Corpus (CroSpeak Corpus) is the corpus of recorded Croatian speech covering radio weather forecasts, radio news, read tales, weather dialogs, and TV news (featuring unspontaneous and spontaneous speech). Overall size of the corpus is 19.35 hours and 227,280 tokens. All utterances have been transcribed following standard Croatian orthography. The corpus has been compiled and processed at Department of Infirmation Sciences, University of Rijeka
relevantPublications	Martinčić-Ipšić, S., Pobar, M., Ipšić, I. Croatian Large Vocabulary Automatic Speech Recognition, Automatika, Vol 52, no 2 (2011), p. 147-157. Martinčić-Ipšić, S., Ipšić, I. Recognition of Croatian Broadcast Speech. Budin, L. (ed.), Ribarić, S., (ed.). XXVII. MIPRO 2004, Opatija, Vol. CTS + CIS, p. 111-114. 2004.
resourceType	corpus
mediaType	speech
lingualityType	monolingual
languageId	hrv
size	227 280
sizeUnit	token

resourceName	Slovene Web Corpus
resourceShortName	slWaC
downloadLocation	http://www.nljubesic.net/projects/slWaC.html
dateCreation	2011

projectPartner	FFZG
iprHolder.organizationName	FFZG
contact.Person.surname	Ljubešić
contact.Person.givenName	Nikola
contact.Person.email	nljubesi@ffzg.hr
DistributionInfo	available-restricted
license	CC BY-SA
resourceLocation	http://www.nljubesic.net/projects/slWaC.html
distributionAccessMedium	downloadable
restrictionsOfUse	following CC BY-SA license restrictions
licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
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 yielding ca 380 million tokens. The corpus has been lemmatised and MSD-tagged automatically using ToTaLe system by Tomaž Erjavec. The corpus is distributed under the CC-BY-SA licence.
relevantPublications	Ljubešić, N., Erjavec, T. (2011) hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene // Proceedings of the 14th International Conference Text, Speech and Dialogue (TSD2011), Plzeň, Czech Republic, 1-5 September 2011, Lecture Notes in Artificial Intelligence 6836, Springer, Heidelberg, pp 395-402.
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	slo
size	380 000 000
sizeUnit	token

resourceName	Serbian Web Corpus
resourceShortName	srWaC
downloadLocation	http://www.nljubesic.net/projects/srWaC.html
dateCreation	ongoing work
projectPartner	FFZG
iprHolder.organizationName	FFZG
contact.Person.surname	Ljubešić
contact.Person.givenName	Nikola
contact.Person.email	nljubesi@ffzg.hr
DistributionInfo	available-restricted
license	CC BY-SA
resourceLocation	http://www.nljubesic.net/projects/srWaC.html
distributionAccessMedium	downloadable
restrictionsOfUse	following CC-BY-SA licence restrictions

licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
description	Serbian Web Corpus (srWaC) is the first version of the Serbian web corpus. It is the work in progress with more than 1 billion of tokens expected. So far 400 million has been collected using the same methodology and tools used for collecting hrWaC and slWaC. The desired size is expected to be reached in 2012-09.
relevantPublications	Ljubešić, N., Erjavec, T. (2011) hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene // Proceedings of the 14th International Conference Text, Speech and Dialogue (TSD2011), Plzeň, Czech Republic, 1-5 September 2011, Lecture Notes in Artificial Intelligence 6836, Springer, Heidelberg, pp 395-402.
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	srp
size	1 000 000 000
sizeUnit	token

resourceName	SouthEast European Parallel Corpus
resourceShortName	SETimes corpus
downloadLocation	http://www.nljubesic.net/resources/corpora/setimes/
dateCreation	2012
projectPartner	FFZG
iprHolder.organizationName	FFZG
contact.Person.surname	Ljubešić
contact.Person.givenName	Nikola
contact.Person.email	nljubesi@ffzg.hr
DistributionInfo	available-restricted use
license	CC BY-SA
resourceLocation	http://www.nljubesic.net/resources/corpora/setimes/
distributionAccessMedium	download
restrictionsOfUse	following CC-BY-SA licence restrictions
licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications

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: Bulgarian, Bosnian, Greek, English, Croatian, Macedonian, Romanian, Albanian and Serbian. 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 following procedures were applied to resolve existing issues: (1) stricter extraction process – no HTML residues present; (2) language identification on every non-English document – non-English online documents contain English material in case the article was not translated into that language; (3) resolving encoding issues in Croatian and Serbian – diacritics were partially lost due to encoding errors – text was rediacritized. The sentence-aligned language combinations are freely downloadable in TMX or TXT/Moses format.
relevantPublications	Tyers, F. M., Serdar Alperen, M. South-East European Times: A parallel corpus of Balkan languages, LREC2010.
resourceType	corpus
mediaType	text
lingualityType	parallel
languageId	alb, bos, bul, eng, gre, hrv, mac, rum, srp, tur
size	43 142 458
sizeUnit	token

resourceName	Croatian-English Giza++ Table
resourceShortName	hrenGiza++
downloadLocation	http://hnk.ffzg.hr/hrenGiza
dateCreation	2012
projectPartner	FFZG
iprHolder.organizationName	FFZG
contact.Person.surname	Agić
contact.Person.givenName	Željko
contact.Person.email	zagic@ffzg.hr
DistributionInfo	available-restricted use
license	CC BY-SA
resourceLocation	http://hnk.ffzg.hr/hrenGiza
distributionAccessMedium	download
restrictionsOfUse	following CC-BY-SA licence restrictions
licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
description	Croatian-English Giza++ phrase table is the Croatian-English translation model built on the basis of several Croatian-English parallel corpora: Croatian-English Parallel Corpus, Croatian-English Parallel Web Corpus (hrenWaC), Croatian-English aligned sentences from SouthEast European Parallel Corpus (SETimes Corpus v2.0).

relevantPublications	
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	parallel
languageId	eng, hrv
size	1 160 274
sizeUnit	entry

resourceName	Web Content Extractor
resourceShortName	Web Content Extractor
downloadLocation	http://www.nljubesic.net/resources/tools/webcontentextractor/
dateCreation	2011
projectPartner	FFZG
iprHolder.organizationName	Nikola Ljubešić
contact.Person.surname	Ljubešić
contact.Person.givenName	Nikola
contact.Person.email	nljubesi@ffzg.hr
DistributionInfo	available-restricted use
license	Apache 2.0
resourceLocation	http://www.nljubesic.net/resources/tools/webcontentextractor/
distributionAccessMedium	downloadable
restrictionsOfUse	following the Apache 2.0 licence
licenseSignatory.Person.position	
foreseenUse	human use, NLP applications
actualUse	human use, NLP applications
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 hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. An implementation (a java file) is published under the Apache 2.0 licence can be downloaded from http://www.nljubesic.net/upload/ContentExtractor.java. It requires jsoup-1.4.1.jar. A Croatian evaluation sample used in the paper can be downloaded from http://www.nljubesic.net/upload/gold_standard.zip and it is distributed under the CC-BY-SA license.
relevantPublications	Ljubešić, N., Erjavec, T. (2011) hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene // Proceedings of the 14th International Conference Text, Speech and Dialogue (TSD2011), Plzeň, Czech Republic, 1-5 September 2011, Lecture Notes in Artificial Intelligence 6836, Springer, Heidelberg, pp 395-402.
resourceType	tool
mediaType	text
lingualityType	language independent
languageId	
size	–
sizeUnit	–

resourceName	Croatian Academic Spelling Checker
resourceShortName	Hascheck
downloadLocation	http://hacheck.tel.fer.hr/
dateCreation	1994
projectPartner	FFZG
iprHolder.organizationName	FER
contact.Person.surname	Dembitz
contact.Person.givenName	Šandor
contact.Person.email	sandor.dembitz@fer.hr
DistributionInfo	available-unrestricted use
license	
resourceLocation	http://hacheck.tel.fer.hr/
distributionAccessMedium	web service
restrictionsOfUse	following the license restrictions
licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
description	Croatian Academic Spelling Checker (Hascheck) is one of the oldest Internet services in Croatia. In various forms it acts as a public service and free spelling checker for text written in Croatian language since the spring of 1994. Hascheck's dictionary database is organized into three sections: (1) Croatian general lexicon, (2) Croatian lexicon of names, (3) English general lexicon. Dictionary database is not static and it is being constantly improved. In the background a working expert system that learns new words from texts submitted for processing. The database is maintained through supervised learning process and currently it exceeds one million types, which all have a been attested in the texts written in Croatian language. The core solutions for spelling checker were written by Šandor Dembitz and the majority of web site processing was written by Gordon Gledec, while the web interface was written by Hrvoje Miholić.
relevantPublications	Dembitz, Š., Knežević, P., Sokele, M. Developing a Spell Checker as an Expert System. // CIT. Journal of computing and information technology. 11 (2003) , 4; 285-292.
resourceType	technologyToolService
mediaType	text
lingualityType	multilingual
languageId	hrv, eng
size	-
sizeUnit	-

5.3. Hungarian language resources detailed specification

resourceName	Hungarian National Corpus
resourceShortName	HNC
downloadLocation	corpus.nytud.hu/hnc
dateCreation	1998-2003
projectPartner	HASRIL
iprHolder.organizationName	HASRIL
contact.Person.surname	Váradi
contact.Person.givenName	Tamás
contact.Person.email	varadi.tamas@nytud.mta.hu
DistributionInfo	avaiable-restricted use
license	CC BY NC SA
resourceLocation	HASRIL
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Deputy director
foreseenUse	human use nlpApplications
actualUse	human use
description	The national corpus of Hungarian language which is derived into five subcorpora by regional language variant and into five subcorpora by text genre. The subcorpus to be studied can be chosen by any combination of these. That makes the HNC an appropriate tool to study the differences not just between text genres but between language variants. HNC aims to be a representative general corpus of present-day standard Hungarian.
relevantPublications	Váradi, Tamás: The Hungarian National Corpus. In: Proceedings of the 3rd LREC Conference, Las Palmas, Spanyolország, 2002, 385-389. http://corpus.nytud.hu/mnsz Sass, Bálint: The Verb Argument Browser. In: Sojka, P. et al. (eds.): Proceedings of TSD 2008, Brno, Czech Republic, 2008, LNCS 5246, 187-192. http://corpus.nytud.hu/vab
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	hu
size	187,000.000
sizeUnit	token

resourceName	Child Language Corpus
resourceShortName	CHILC
downloadLocation	--
dateCreation	2012
projectPartner	HASRIL
iprHolder.organizationName	HASRIL
contact.Person.surname	Pintér
contact.Person.givenName	Tibor
contact.Person.email	pinter.tibor@nytud.mta.hu
DistributionInfo	avaiable-restricted use
license	
resourceLocation	HASRIL
distributionAccessMedium	other
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Kinga Mátyus
foreseenUse	human use
actualUse	human use
description	The child language corpus consists of 60 interviews with 4/6-5/6 year-old Hungarian children (from Budapest and from different socio-economical backgrounds), with more than 30 hours recording. The interviews include several tasks (picture-based story-telling, telling the rules of a well-known game) and guided conversation. Each interview was conducted by an adult tester. The resource is available in chat (CHILDES) transcription format.
relevantPublications	-
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	hu
size	60
sizeUnit	other

resourceName	Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection
resourceShortName	ParallelSpeech-hu
downloadLocation	
dateCreation	2009-07-01 - 2012-09-30
projectPartner	BME
iprHolder.organizationName	Budapest University of Technology and Economics
contact.Person.surname	Németh

contact.Person.givenName	Géza
contact.Person.email	nemeth@tmit.bme.hu
DistributionInfo	avaiable-restricted use
license	CLARIN_RES
resourceLocation	BME-TMIT
distributionAccessMedium	DVD
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	head of dept.
foreseenUse	human use nlpApplications
actualUse	human use nlpApplications
description	This speech database contains 2,000 sentences. Each speaker reads this sentence set. This parallel speech database is used to train HMM based TTS and for unit selection TTS.
relevantPublications	
resourceType	corpus
mediaType	text audio
lingualityType	monolingual
languageId	hu
size	25,5
sizeUnit	hour

resourceName	Read speech database in Hungarian
resourceShortName	ReadSpeech-hu
downloadLocation	
dateCreation	2005-01-01 - 2012-07-10
projectPartner	BME
iprHolder.organizationName	Budapest University of Technology and Economics
contact.Person.surname	Németh
contact.Person.givenName	Géza
contact.Person.email	nemeth@tmit.bme.hu
DistributionInfo	avaiable-restricted use
license	CLARIN_RES
resourceLocation	BME-TMIT
distributionAccessMedium	DVD
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	head of dept.
foreseenUse	human use nlpApplications
actualUse	human use nlpApplications

description	The first automatic TTS based Hungarian weather forecast application (www.metnet.hu) based on this database.
relevantPublications	http://www.springerlink.com/content/mr6m71133887823m
resourceType	corpus
mediaType	text audio
lingualityType	monolingual
languageId	hu
size	10
sizeUnit	hour

resourceName	Di-phone database for text-to-speech conversion
resourceShortName	Di-phone-hu
downloadLocation	http://speechlab.tmit.bme.hu/CESAR/diphone_hu.zip
dateCreation	2007-01-01- 2011-08-20
projectPartner	BME
iprHolder.organizationName	Budapest University of Technology and Economics
contact.Person.surname	Németh
contact.Person.givenName	Géza
contact.Person.email	nemeth@tmit.bme.hu
DistributionInfo	avaiable-restricted use
license	CLARIN_RES
resourceLocation	Not available yet
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	head of dept.
foreseenUse	human use nlpApplications
actualUse	human use nlpApplications
description	The Profivox hungarian di-phone TTS uses a database based on this resource.
relevantPublications	Olaszy G. - Gépi beszédkeltés információs rendszerekhez Magyarországon. AKUSZTIKAI SZEMLE III:(1-3) pp. 4-13. (1999)
resourceType	corpus
mediaType	text audio
lingualityType	monolingual
languageId	hu
size	1,646
sizeUnit	second

resourceName	Hungarian BABEL phonetic and prosodic segmentation and syntactic analysis
resourceShortName	BABEL-hu-Addon1

downloadLocation	if applicable
dateCreation	2012-03-31
projectPartner	BME-TMIT
iprHolder.organizationName	Budapest University of Technology and Economics, Dept of Telecommunications and Media Informatics
contact.Person.surname	Szaszák
contact.Person.givenName	György
contact.Person.email	szaszak@tmit.bme.hu
DistributionInfo	please, choose one of the values available-unrestricted use avaiable-restricted use notAvailable underNegotiation
license	META-SHARE NR NC
resourceLocation	BME-TMIT
distributionAccessMedium	CD-ROM or downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Head of dept.
foreseenUse	please, leave the appropriate human use nlpApplications
actualUse	please, leave the appropriate human use nlpApplications
description	This resource is an add-on to Hungarian BABEL speech corpus, containing phoneme level segmentation, prosodic segmentation and annotation for phonological phrases and disambiguated syntactic analysis for 330 sentences/utterances from Hungarian BABEL.
relevantPublications	Szaszak Gyorgy, Nagy Katalin, Beke Andras: Analysing the correspondence between automatic prosodic segmentation and syntactic structure. In: INTERSPEECH-2011 Conference Proceedings: Speech Science and Technology for Real Life. Florence, Italy 2011.08.27-2011.08.31. ISCA, pp. 1057-1060.
resourceType	corpus
mediaType	text audio
lingualityType	monolingual
languageId	hu
size	330
sizeUnit	sentence utterance

resourceName	Hungarian Spontaneous Speech Database
resourceShortName	BEA
downloadLocation	--
dateCreation	2012
projectPartner	HASRIL
iprHolder.organizationName	HASRIL

contact.Person.surname	Pintér
contact.Person.givenName	Tibor
contact.Person.email	pinter.tibor@nytud.mta.hu
DistributionInfo	avaiable-restricted use
license	CC BY NC SA
resourceLocation	HASRIL
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse commercialUse
licenseSignatory.Person.position	Mária Gósy
foreseenUse	human use
actualUse	human use
description	BEA is a multi-functional speech database of Hungarian that contains various types of spontaneous speech (including conversations) and sentence repetitions and reading. This is the largest speech database of Hungarian consisting of about 270 hour recorded speech material of 265 speakers at present. The number of annotated materials is close to 50%. The recording circumstances of the speech materials are constant, showing a high technical background. All the speakers are recorded in the same sound attenuated room of the Research Institute for Linguistics (HAS, Budapest) that makes it possible to analyse the spontaneous speech samples from various acoustic-phonetic and linguistic aspects. In addition, the BEA Database provides a unique possibility also for the research of speech technology. The speech materials of BEA are annotated at various levels of transcription. The current version contains 135 hours of recordings together with the transcribed (in transcriber) and aligned text of 179 speakers.
relevantPublications	
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	hu
size	270
sizeUnit	hour

resourceName	Tesztel Hungarian Noisy Telephone Speech Corpus
resourceShortName	Not applicable
downloadLocation	if applicable
dateCreation	2006-12-31
projectPartner	BME-TMIT
iprHolder.organizationName	BME-TMIT
contact.Person.surname	Szaszák

contact.Person.givenName	György
contact.Person.email	szaszak@tmit.bme.hu
DistributionInfo	notAvailable
license	notAvailable
resourceLocation	BME-TMIT
distributionAccessMedium	DVD-R
restrictionsOfUse	informResourceOwner academic-nonCommercialUse attribution shareAlike
licenseSignatory.Person.position	Head of dept.
foreseenUse	human use nlpApplications
actualUse	human use nlpApplications
description	This database contains noisy speech samples for noise robust ASR in Hungarian, recorded via PSTN and mobile telephone network. The aim of this project was to create a mobile phone voice based Hungarian speech database recorded in noisy environments for testing purposes (also called Tesztel). The database contains voices of 100 speakers, recorded through mobile telephone in noisy environments. The main goal was to test phoneme based recognizers, which have been already trained, so the corpus must have been compact and had to cover as good as possible the specific character of the Hungarian language. The text that the speaker had to tell was designed to contain at least one of every Hungarian phoneme, taking in consideration the statistics of phonemes, diphones, triphones and syllables in Hungarian language. The corpus contains not only continuously told sentences, but command words, spelled forenames, numbers, dates, different currency types, city names, questions with yes/no answer, phonetically rich words. The database contains mostly spontaneous speech. Since the whole database contains speech recorded in noisy environments, we wanted to find out an average value of the signal to noise ratio for the recorded speech. But this parameter depends on multiple and different factors, such as the type and intensity of the background noise and the parameters of the channel. Probably, this is the reason why the measured signal to noise ratio varies on a very large scale, between 5dB and 25dB. The lowest value (app. 5dB) was measured at the recordings that were made near busy highways or on public transport (mainly old trams) in the rush hour. The highest values (app. 25dB) were measured at the recordings that were made on side streets, public transport or room (mainly late at night).
relevantPublications	
resourceType	corpus
MediaType	audio
lingualityType	monolingual
languageId	hu
size	100
sizeUnit	speaker

resourceName	Hungarian Child Database for Speech Processing Applications
resourceShortName	Not applicable
downloadLocation	if applicable
dateCreation	
projectPartner	BME-TMIT
iprHolder.organizationName	BME-TMIT
contact.Person.surname	Szaszák
contact.Person.givenName	György
contact.Person.email	szaszak@tmit.bme.hu
DistributionInfo	notAvailable
license	notAvailable
resourceLocation	BME-TMIT
distributionAccessMedium	DVD-R
restrictionsOfUse	informResourceOwner academic-nonCommercialUse attribution shareAlike
licenseSignatory.Person.position	Head of dept.
foreseenUse	human use nlpApplications
actualUse	human use nlpApplications
description	Child speech for Hungarian, useful for speech training and language learning applicatzions for children. More detailed description of the corpus was uploaded to: http://alpha.tmit.bme.hu/speech/paperc013.php
relevantPublications	Vicsi, K. - Csatári, F. - Bakcsi, Zs. "Distance Score evaluation of the visualized speech spectra at audio-visual articulation training" - EUROSPEECH. Vicsi, K. - Roach, P. - Öster, A. - Kacic, Z. - Barczikay, P. - Sinka, I. : SPECO, a multimedia multilingual teaching and training system for speech handicapped children EUROSPEECH '99
resourceType	corpus
MediaType	audio
lingualityType	monolingual
languageId	hu
size	72
sizeUnit	speaker

resourceName	BABEL and MRBA sentence modality annotation for Hungarian
resourceShortName	Not applicable
downloadLocation	if applicable
dateCreation	2008-12-31
projectPartner	BME-TMIT
iprHolder.organizationName	BME-TMIT
contact.Person.surname	Szaszák

contact.Person.givenName	György
contact.Person.email	szaszak@tmit.bme.hu
DistributionInfo	notAvailable
license	notAvailable
resourceLocation	BME-TMIT
distributionAccessMedium	DVD-R
restrictionsOfUse	informResourceOwner academic-nonCommercialUse attribution shareAlike
licenseSignatory.Person.position	Head of dept.
foreseenUse	human use nlpApplications
actualUse	human use NlpApplications
description	Corpus holding modality annotations for subparts of Hungarian BABEL and MRBA corpora. If extended, a useful source for research and analysis or recognition of sentence modality.
relevantPublications	Vicsi K, Szaszák Gy: Using prosody to improve automatic speech recognition. SPEECH COMMUNICATION 52:(5) pp. 413-426. (2010)
resourceType	corpus
MediaType	audio
lingualityType	monolingual
languageId	hu
size	50
sizeUnit	speaker

resourceName	Hungarian Human-Computer Interaction Technologies Multimodal Database
resourceShortName	HUCOMTECH
downloadLocation	–
dateCreation	2011
projectPartner	HASRIL
iprHolder.organizationName	University of Debrecen
contact.Person.surname	Pintér
contact.Person.givenName	Tibor
contact.Person.email	pinter.tibor@nytud.mta.hu
DistributionInfo	underNegotiation
license	--
resourceLocation	University of Debrecen
distributionAccessMedium	accessibleThroughInterface

restrictionsOfUse	Other
licenseSignatory.Person.position	László Hunyadi
foreseenUse	human use, nlpApplications
actualUse	human use, nlpApplications
description	The first and only resource for Hungarian language of aligned text-video-audio segments. The alignment is made by speech units. The audio-visual database recording and annotation project is carried out by the HuComTech (Hungarian Human-Computer Interaction Technologies) multidisciplinary research team, involving computer scientists (for digital image processing), computational linguists (for speech processing) and communication experts (between June 2009 and June 2011). The HuComTech (Hungarian Human-Computer Interaction Technologies) project aims to build a multimodal (audio and visual) database of Hungarian language. The research group aims to contribute to the knowledge of the interplay of prosodic, verbal and nonverbal features of communicative events by the examination of annotated spontaneous dialogues. Both, formal guided job interviews and informal semi-guided conversations have been recorded with approximately 110 Hungarian university students, resulting in a huge Hungarian audio-visual database complemented with detailed, multi-level multimodal annotation.
relevantPublications	Hunyadi, L. 2012. Collaboration in Virtual Space in Digital Humanities. In: Deegan, M., MacCarty, W. (eds.): Collaborative Research in the Digital Humanities. Ashgate. 93-103. Hunyadi, L. 2011. Multimodal Human Computer Interaction Technologies. Theoretical Modeling and Application in Speech Processing. Argumentum 7. 240 - 260. Staudt A., Pápay K. 2011. The Annotation of the HuComTech Audio Database in Practice, Observations and Questions Arising. Argumentum 7. 313-329.
resourceType	corpus
mediaType	text audio video
lingualityType	monolingual
languageId	hu
size	50
sizeUnit	hour

resourceName	ht-online
resourceShortName	ht-online
downloadLocation	Ht.nytudhu/htonline
dateCreation	2009-2012
projectPartner	HASRIL
iprHolder.organizationName	Termini Research Centre

contact.Person.surname	Pintér
contact.Person.givenName	Tibor
contact.Person.email	pinter.tibor@nytud.mta.hu
DistributionInfo	available-unrestricted use
license	
resourceLocation	HASRIL
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	János Péntek
foreseenUse	nlpApplications
actualUse	human use
description	A unique lexical database of the most common loanwords in Hungarian language used outside Hungary (collected from 7 regions). The database should be used as special lexical resource in the Hungarian language tools based on the Hungarian morphology.
relevantPublications	-
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	monolingual
languageId	hu
size	4,000
sizeUnit	entry

resourceName	Hungarian Concise Dictionary (with sample sentences)
resourceShortName	HCD
downloadLocation	–
dateCreation	2011
projectPartner	HASRIL
iprHolder.organizationName	TINTA Publishing House
contact.Person.surname	Pintér
contact.Person.givenName	Tibor
contact.Person.email	pinter.tibor@nytud.mta.hu
DistributionInfo	underNegotiation
license	--
resourceLocation	HASRIL
distributionAccessMedium	hardDisk
restrictionsOfUse	other

licenseSignatory.Person.position	Gábor Kiss
foreseenUse	please, leave the appropriate human use nlpApplications
actualUse	please, leave the appropriate human use nlpApplications
description	A unique dictionary of Hungarian language of 16 000 headwords (entries) followed by frequency data. Each entry describes the most common forms (given by pragmatical reasons) of the headword. The entries are divided into meanings which counts 33 000 carefully selected and stylistically labelled meanings. The dictionary contains sentences brought from real language use and 3000 phrasems.
relevantPublications	--
resourceType	lexicalConceptualResource
mediaType	txt
lingualityType	monolingual
languageId	hu
size	16,000
sizeUnit	entry

resourceName	High-Speed Unification Morphology
resourceShortName	HUMor
downloadLocation	--
dateCreation	1991 –
projectPartner	HASRIL
iprHolder.organizationName	MorphoLogic Ltd.
contact.Person.surname	Pintér
contact.Person.givenName	Tibor
contact.Person.email	pinter.tibor@nytud.mta.hu
DistributionInfo	underNegotiation
license	--
resourceLocation	--
distributionAccessMedium	downloadable
restrictionsOfUse	other
licenseSignatory.Person.position	Gábor Prószéky
foreseenUse	nlpApplications
actualUse	nlpApplications

description	Humor, a reversible, string-based unification approach for lemmatizing and disambiguating language data, has been used for both language corpus analysis and creation of a variety of linguistic software applications such as spell-checking. The system is language-independent, allowing multilingual applications for a variety of language types. Its Hungarian version, the largest and most precise implementation, contains nearly 100,000 stems. The system has been tested rigorously by both, linguists and end-users of word-processing tools. Humor-based linguistic modules have been licensed by major software producers, and the lemmatizer has been used in lexicographic research since 1991. One tool provides disambiguation, tagging, and parsing functions. The system can describe various natural languages, including both Eastern European and non-Eastern European languages. Several Humor subsystems for different purposes (lemmatizing, hyphenating, spell-checking/correcting, grammar checking) are commercially available, and have been built into several major word-processing and full-text retrieval systems. An inflectional thesaurus and a series of intelligent bilingual dictionaries have also been developed. (MSE)
relevantPublications	Gábor Prószéky 1995. Humor (High-Speed Unification Morphology): A Morphological System for Corpus Analysis. In: Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995); see FL 024 759.
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	hu
size	100,000
sizeUnit	entry

resourceName	Graphical Query Interface for Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection
resourceShortName	
downloadLocation	if applicable
dateCreation	Under creation
projectPartner	BME-TMIT
iprHolder.organizationName	University of Debrecen
contact.Person.surname	Olaszy
contact.Person.givenName	Gábor
contact.Person.email	olaszy@tmit.bme.hu
DistributionInfo	underNegotiation
license	underNegotiation
resourceLocation	underNegotiation
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Head of dept.

foreseenUse	human use nlpApplications
actualUse	Resource under creation
description	This interface will be designed to allow for fast access, searching and statistical query possibilities and functionality for the Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection in order to allow for advanced phonetic/speech technology research on the corpus.
relevantPublications	
resourceType	technologyToolService
mediaType	text audio
lingualityType	monolingual
languageId	hu
size	Unknown yet
sizeUnit	other

resourceName	Multilingual Speech Segmentation Tool
resourceShortName	Not applicable
downloadLocation	If applicable
dateCreation	2012-12-31
projectPartner	BME-TMIT
iprHolder.organizationName	BME-TMIT
contact.Person.surname	Szaszák
contact.Person.givenName	György
contact.Person.email	szaszak@tmit.bme.hu
DistributionInfo	notAvailable
license	notAvailable
resourceLocation	BME-TMIT
distributionAccessMedium	DVD-R
restrictionsOfUse	informResourceOwner academic-nonCommercialUse attribution noDerivetives
licenseSignatory.Person.position	Head of dept.
foreseenUse	human use nlpApplications
actualUse	human use nlpApplications
description	A multilingual speech segmentation tool, capable of phoneme segmentation of utterances for 6 languages: English, French, Italian, Spanish and Hungarian.
relevantPublications	
resourceType	technologyToolService
MediaType	audio
lingualityType	monolingual
languageId	hu
size	
sizeUnit	other

resourceName	Sentence Modality Recognizer
resourceShortName	Not applicable
downloadLocation	if applicable
dateCreation	2008-12-31
projectPartner	BME-TMIT
iprHolder.organizationName	BME-TMIT
contact.Person.surname	Szaszák
contact.Person.givenName	György
contact.Person.email	szaszak@tmit.bme.hu
DistributionInfo	notAvailable
license	notAvailable
resourceLocation	BME-TMIT
distributionAccessMedium	DVD-R
restrictionsOfUse	informResourceOwner academic-nonCommercialUse attribution shareAlike
licenseSignatory.Person.position	Head of dept.
foreseenUse	human use nlpApplications
actualUse	human use nlpApplications
description	Sentence modality recognizer from speech for Hungarian and German, can be used in speech recognition and understanding.
relevantPublications	Vicsi K, Szaszák Gy: Using prosody to improve automatic speech recognition. SPEECH COMMUNICATION 52:(5) pp. 413-426. (2010)
resourceType	technologyToolService
MediaType	audio
lingualityType	bilingual
languageId	hu
size	
sizeUnit	other

5.4. **Polish** language resources detailed specification

resourceName	Składnica
resourceShortName	Składnica
downloadLocation	http://zil.ipipan.waw.pl/Składnica
dateCreation	2008-2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Woliński
contact.Person.givenName	Marcin
contact.Person.email	marcin.wolinski@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	GPL3
resourceLocation	http://zil.ipipan.waw.pl/Składnica
distributionAccessMedium	downloadable
restrictionsOfUse	attribution, shareAlike
foreseenUse	NlpApplications
actualUse	NlpApplications
description	Składnica is the result of the Polish Ministry of Science and Higher Education research grant (ended in October 2011) on construction of a treebank for Polish using automatic syntactic analysis. The resource is a treebank of Polish constituents created automatically and then manually corrected.
relevantPublications	not yet available
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	pl
size	8,227
sizeUnit	sentence

resourceName	The Corpus of Polish Summaries
resourceShortName	SummaryCorpus
downloadLocation	–
dateCreation	2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Ogrodniczuk
contact.Person.givenName	Maciej
contact.Person.email	maciej.ogrodniczuk@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	to be defined
resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	nlpApplications
actualUse	nlpApplications

description	The corpus of summaries is going to have human-written summaries of 154 texts, each text sized between 1000 and 4000 words. These texts were extracted from "Rzeczpospolita" Corpus (http://www.cs.put.poznan.pl/dweiss/rzeczpospolita) – a corpus of press articles from the website of the "Rzeczpospolita" newspaper. The set of articles contains articles published since year 1993 to 2002 and is yet not freely available. A set of frequently represented text categories in the "Rzeczpospolita" Corpus was chosen: economics, law, news from Poland, culture, sport, science and technology, opinions. The corpus will contain two types of summaries: abstractive and extractive. Each text is going to have 3 summaries of both types, varying in length: a 20%, 10% and 5% summary (in terms of word count of original text). Abstractive summaries are simply written by annotators as a free text, extractive summaries are created by selecting unconstrained fragments of the original text (however following some guidelines) in terms of single character as the smallest possible selection. The 10% extractive summary contains only a subset of selections in 20% extractive summary etc.
relevantPublications	not yet available
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	pl
size	yet unknown
sizeUnit	–

resourceName	Parallel English-Polish Corpus
resourceShortName	ParallelCorpus
downloadLocation	–
dateCreation	2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Ogrodniczuk
contact.Person.givenName	Maciej
contact.Person.email	maciej.ogrodniczuk@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	to be defined
resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	nlpApplications
actualUse	nlpApplications
description	Parallel corpus of texts from various domains.
relevantPublications	not yet available
resourceType	corpus
mediaType	text
lingualityType	bilingual
languageId	pl, eng
size	3,000,000
sizeUnit	words each side

resourceName	Redistributable Polish-Russian Corpus
resourceShortName	DistrPLRU
downloadLocation	–
dateCreation	2011-2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Ogrodniczuk
contact.Person.givenName	Maciej
contact.Person.email	maciej.ogrodniczuk@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	to be defined
resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	NlpApplications
actualUse	NlpApplications
description	Polish-Russian Parallel Corpus is currently being created to reach 50 million words, 50% of Polish originals translated into Russian and 50% vice versa. The core of the resources consists of the literary classics of the nineteenth century and contemporary works which are most popular in the neighbouring country. The corpus contains press texts and their translations, as well as legal texts. The texts are annotated according to the DTDs of the National Corpus of Polish and the Russian National Corpus. A morphosyntactic search is possible, although the standards of the two national corpora differ according to some grammatical classes and categories.
relevantPublications	not yet available
resourceType	corpus
mediaType	text
lingualityType	bilingual
languageId	pl, rus
size	yet unknown
sizeUnit	–

resourceName	Learner Speech Database
resourceShortName	PESLC
downloadLocation	To be determined
dateCreation	not applicable
projectPartner	University of Łódź
iprHolder.organizationName	University of Łódź
contact.Person.surname	Pęzik
contact.Person.givenName	Piotr
contact.Person.email	contact@pelcra.pl
DistributionInfo	
license	CC-BY-NC
resourceLocation	To be determined
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse

licenseSignatory.Person.position	Associate Professor
foreseenUse	human use nlpApplications
actualUse	human use nlpApplications
description	The Learner Speech Database contains samples of spoken learner English from the PELCRA Learner English Corpus (PLEC). The database contains transcriptions of Poles speaking in English and in Polish on a variety of informal topics. The transcriptions are time-aligned at the level of utterances with the underlying recordings, most of which are studio-quality and uncompressed. Possible NlpApplications of this database include the improvement of speech recognition systems dedicated for speakers of English with a Polish accent.
relevantPublications	1. Pęzik P. 2012 Towards the PELCRA Learner English Corpus. In Corpus Data across Languages and Disciplines, Lodz Studies in Language. Vol. 28. Edited by Piotr Pęzik. Peter Lang. Frankfurt am Main. Forthcoming. 2. Molenda M., Pęzik P. 2012 A corpus-based study of learners' confluence. In TaLC 10. Forthcoming. 3. Zając M, Pęzik P. 2012 Annotating pronunciation errors in the PLEC spoken learner corpus. In TaLC 10. Forthcoming.
resourceType	corpus
mediaType	text audio
lingualityType	monolingual
languageId	en
size	
sizeUnit	word
size	50,000
sizeUnit	word

resourceName	SNUV Voice Recognition Speech Database
resourceShortName	SNUV
downloadLocation	not applicable
dateCreation	not applicable
projectPartner	University of Łódź
iprHolder.organizationName	Voice Lab sp. z o.o.
contact.Person.surname	Szwelnik
contact.Person.givenName	Tomasz
contact.Person.email	tomasz.szwelnik@voicelab.pl
DistributionInfo	underNegotiation
license	CC-BY
resourceLocation	To be determined
distributionAccessMedium	downloadable

restrictionsOfUse	academic-nonCommercialUse commercialUse
licenseSignatory.Person.position	Assistant Professor
foreseenUse	NlpApplications
actualUse	NlpApplications
description	SNUV is a spelling and number and recognition speech database composed of 200 hours of recordings of Polish speakers reading numbers and spelling words, recorded in 22050kHz, 16-bit *.wav files. It was developed in a large-scale crowd-sourcing includes a transcription of the recordings in text format, encoded in the UTF-8 standard. The purpose of this resource is to enable the creation of automatic speech recognition (ASR) tools that allow the user to spell out a word or a number to be recognize.. SNUV is potentially the largest available Polish speech recognition database, which can be released under a CC-license.
relevantPublications	To be determined
resourceType	corpus
mediaType	audio text
lingualityType	monolingual
languageId	pl
size	200
sizeUnit	hour

resourceName	PELCRA Time-Aligned Spoken Corpus
resourceShortName	TASC
downloadLocation	http://pelcra.pl/resources/spoken/pelcra_sp_2.tgz.
dateCreation	2012
projectPartner	University of Łódź
iprHolder.organizationName	University of Łódź
contact.Person.surname	Piotr
contact.Person.givenName	Pęzik
contact.Person.email	piotr.pezik@uni.lodz.pl
DistributionInfo	available, unrestricted use
license	CC-BY-NC license
resourceLocation	–
distributionAccessMedium	downloadable
restrictionsOfUse	
foreseenUse	nlpApplications
actualUse	nlpApplications

description	The corpus is the largest collection of transcriptions of naturally occurring conversational Polish has been compiled by the PELCRA team at the University of Łódź since 2000, initially as part of the PELCRA Reference Corpus and later within the National Corpus of Polish. The corpus contains over 43 hours of conversation recorded in an informal setting. So far, this data has been only available through online search interfaces, but within the CESAR project a subset has been made available in the TEI P5 format. The transcriptions have been time-aligned with the original recordings at the level of utterances and made available under the CC-BY-NC license.
relevantPublications	Piotr Pęzik 2012 Język mówiony w NKJP. In Narodowy Korpus Języka Polskiego. Wydawnictwo Naukowe PWN, Warsaw. 2012.]
resourceType	corpus
mediaType	text, audio
lingualityType	monolingual
languageId	pl
size	40
sizeUnit	hour

resourceName	Paralela DB
resourceShortName	Paralela
downloadLocation	http://pelcra.pl/paralela
dateCreation	2012
projectPartner	University of Łódź
iprHolder.organizationName	University of Łódź
contact.Person.surname	Piotr
contact.Person.givenName	Pęzik
contact.Person.email	piotr.pezik@uni.lodz.pl
DistributionInfo	available, unrestricted use
license	CC-BY-NC license
resourceLocation	–
distributionAccessMedium	downloadable
restrictionsOfUse	
foreseenUse	nlpApplications
actualUse	nlpApplications
description	Paralela database (Paralela DB) is a multilingual parallel corpus containing texts of CORDIS news database, RAPID press release of the EU, press releases of the European Parliament and of the European Southern Observatory. Except for Polish which is obligatory, the database covers more than 20 other languages. The process of converting, processing and exporting parallel resources encoded in a variety of formats (ranging from HTML and PDF to TEI) is facilitated by the use of a central relational database system (named Paralela) to which text collections are imported. The Paralela database is used to store bibliographic, structural and alignment information, and is designed to handle multiple alignments of the same collection. Once the variously encoded collections are converted and normalised, they can be processed and exported into more uniform and standard formats used for the exchange of parallel corpora and translation memories.

relevantPublications	
resourceType	corpus
mediaType	text
lingualityType	multilingual
languageId	pl, bg, en, de, da, nl, ru, sl, sw, tr, no, and many other
size	50,000,000
sizeUnit	word

resourceName	LFG Grammar of Polish
resourceShortName	LFGGrammarPL
downloadLocation	–
dateCreation	2011-2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Przepiórkowski
contact.Person.givenName	Adam
contact.Person.email	adam.przepiorkowski@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	to be defined
resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	nlpApplications
actualUse	nlpApplications
description	New LFG Grammar of Polish is currently being constructed by making extensive reuse of existing language resources for Polish. Its constituent structure (c-structure) is based on a DCG grammar of Polish and the functional structure (f-structure) was mainly inspired by the available HPSG analyses of Polish. Valence information from the dictionary which accompanies the DCG grammar was converted, so that sub categorisation is stated in terms of grammatical functions rather than categories; additionally, missing valence frames may be extracted from the treebank. The obtained grammar will be evaluated using constructed test suites (half of which were provided by previous grammars) and the treebank.
relevantPublications	Patejuk A., Przepiórkowski A. Towards an LFG parser for Polish: An exercise in parasitic grammar development. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pp. 3849-3852, Istanbul, Turkey, 2012. ELRA.
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	monolingual
languageId	pl
size	yet unknown
sizeUnit	–

resourceName	Formal Grammar of Polish
resourceShortName	GFJP
downloadLocation	–
dateCreation	1992-2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Świdziński
contact.Person.givenName	Marek
contact.Person.email	m.r.swidzinski@uw.edu.pl
DistributionInfo	available, unrestricted use
license	to be defined
resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	NlpApplications
actualUse	NlpApplications
description	Formal Grammar of Polish (GFJP) is the most extensive and most detailed formal grammar of Polish expressed as a metamorphosis grammar with several extensions, e.g. allowing for permuting phrases. Syntactic units are represented by terms with parameters formalizing various grammatical features of those units. Rules of the grammar define particular units as sequences of other units and establish correspondences between grammatical features (unification). Agreements are accounted for by parameter matching using an extensive set of parameters. The values a given unit is assigned, be it from the top (“syntactic” features) or from the bottom (“lexical” features), use to spread down the syntactic tree, reaching most of its constituents. Rules defining different syntactic units (sentences or phrases) follow one format. FGP has an ambition to define the whole language – i.e., most structures of Polish are covered.
relevantPublications	not yet available
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	monolingual
languageId	pl
size	460
sizeUnit	rule

resourceName	Syntatic-Gnerative Dctionary of Polish Vrbs
resourceShortName	SSGCP
downloadLocation	–
dateCreation	1980-1995, 2000
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Ogrodniczuk
contact.Person.givenName	Maciej
contact.Person.email	maciej.ogrodniczuk@ipipan.waw.pl
DistributionInfo	available, unrestricted use

license	to be defined
resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	nlpApplications
actualUse	nlpApplications
description	Syntactic-generative dictionary of Polish verbs has been published in paper form in the 1980s-90s. Then its computer implementation in the form of a MS Access database was created. Currently, a better representation format is being constructed for the resource.
relevantPublications	not yet available
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	monolingual
languageId	pl
size	10,559
sizeUnit	verb

resourceName	Polish OpenCYC Lxicon
resourceShortName	OpenCYCPL
downloadLocation	–
dateCreation	2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Pohl
contact.Person.givenName	Aleksander
contact.Person.email	apohllo@o2.pl
DistributionInfo	available, unrestricted use
license	to be defined
resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
licenseSignatory.Person.position	Resource author
foreseenUse	NlpApplications
actualUse	NlpApplications
description	OpenCYC lexicon is currently being translated into Polish and Polish translation are being aligned with English data.
relevantPublications	Pohl A. The semi-automatic construction of the Polish Cyc Lexicon. Investigationes Linguisticae, vol. XXI, s. 17-38, ISSN 1733-1757, 2010.
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	bilingual
languageId	pl, en
size	approx. 20-25,000
sizeUnit	concept

resourceName	Polish-English Wikipedia NE dictionaries
resourceShortName	NERDict
downloadLocation	http://pelcra.pl/res/ecl-dictionaries
dateCreation	2012
projectPartner	University of Łódź
iprHolder.organization Name	University of Łódź
contact.Person.surname	Piotr
contact.Person.givenName	Pęzik
contact.Person.email	piotr.pezik@uni.lodz.pl
DistributionInfo	available, unrestricted use
license	CC
resourceLocation	–
distributionAccessMedium	downloadable
restrictionsOfUse	-
licenseSignatory.Person .position	
foreseenUse	nlpApplications
actualUse	nlpApplications
description	Wikipedia-derived English-Polish and Polish-English thematic dictionaries are based on existing Wikipedia categories, but being manually checked for inappropriately-placed entries. Subjects that are covered include US universities, world cities and villages, Polish artists, journalists, scientists, companies, organisations, etc. The dictionaries are stored in the RDF (Resource Description Framework) program. The categories presented do not reflect the exact Wikipedia structure, but rather conceptual relations.
relevantPublications	
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	bilingual
languageId	pl, en
size	
sizeUnit	

resourceName	Lexeme Forge
resourceShortName	LexemeForge
downloadLocation	–
dateCreation	2011-2012
projectPartner	IPIPAN
iprHolder.organization Name	IPIPAN
contact.Person.surname	Woliński
contact.Person.givenName	Marcin
contact.Person.email	marcin.wolinski@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	to be defined

resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	NlpApplications
actualUse	NlpApplications
description	Lexeme Forge is a Web-based tool used to manage creation of morphological dictionaries for inflectional languages. The system manages a database of lexemes and makes it possible to edit their descriptions, first of all to characterize their inflectional paradigms. The database is modelled after Grammatical Dictionary of Polish, in particular its inflectional patterns are used directly. The system makes it possible to attach various labels to lexemes. Besides typical dictionary labels like informal or dated, special labels are used for excluding some forms from spell-checking dictionaries. This way a special variant of the dictionary can be generated which does not contain certain theoretically correct but extremely infrequent words (i.e., potential false negatives in spell-checking). Moreover, the system makes it possible to specify a classification scheme (or several classification schemes), which the lexemes are to follow. This mechanism is currently used to classify lexemes into common and proper names (with some subclasses).
relevantPublications	See e.g. Woliński M., Miłkowski M., Ogrodniczuk M., Przepiórkowski A. PoliMorf: a (not so) new open morphological dictionary for Polish. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), pp. 860–864. ELRA, Istanbul, Turkey.
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	pl
size	–
sizeUnit	–

resourceName	Slowal
resourceShortName	Slowal
downloadLocation	–
dateCreation	2012
projectPartner	IPIPAN

iprHolder.organization Name	IPIPAN
contact.Person.surname	Skwarski
contact.Person.givenName	Filip
contact.Person.email	filip.skwarski@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	to be defined
resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	nlpApplications
actualUse	nlpApplications
description	Slowal is a Web tool designed for creating valence dictionaries based on the format presented by Filip Skwarski. It describes lemmas by a list of individual frames presented as tables which can be expanded by adding to them new positions, arguments, series of characteristics and examples showing their usage. The tool provides user group management (Guests – add notes to created lemmas, Lexicographers – responsible for expanding existing lemmas, Superlexicographers – responsible for checking correctness of lexicographers work, managing vocabularies and adding new lemmata). Slowal implements a list of features helping in creating and expanding lemmas, e.g.: looking for similar lemmas, validation of created frames, series of filters which can help found lemmas using specific position or arguments and many more. Such created vocabularies can be imported from text format. Slowal is implemented using Django framework.
relevantPublications	not yet available
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	pl
size	yet unknown
sizeUnit	–

resourceName	Lakon
resourceShortName	Lakon
downloadLocation	http://www.cs.put.poznan.pl/dweiss/research/lakon/lakon.zip
dateCreation	2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Dudczak
contact.Person.givenName	Adam
contact.Person.email	maneo@man.poznan.pl
DistributionInfo	available, unrestricted use
license	to be defined
resourceLocation	http://www.cs.put.poznan.pl/dweiss/research/lakon/index_en.html
distributionAccessMedium	Downloadable
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	nlpApplications
actualUse	nlpApplications
description	Lakon is a Polish extractive summarizer using algorithms based on salient sentence selection, namely: heuristic evaluation of position of sentences in paragraphs, word weighting schema tf-idf and okapi bm25 as well as lexical chains combined with thesaurus use. The quality of the automatically generated summaries have been evaluated against a corpus of manually created summaries of selected press articles.
relevantPublications	Dudczak A., Stefanowski J., et al. Automatyczna selekcja zdań dla tekstów prasowych. Institute of Computing Science, Poznań University of Technology, Poland, Research Report RA-03/08, 2008. Dudczak A., Stefanowski J., et al. Comparing Performance of Text Summarization Methods on Polish News Articles. IIPWM'2008 Conference, Zakopane, Poland, 2008. Dudczak A., Stefanowski J., et al. Evaluation of Sentence-Selection Text Summarization Methods on Polish News Articles. Foundations of Computing and Decision Sciences, 1 (vol. 35), 2010, pp. 27—41.
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	pl
size	yet unknown
sizeUnit	–

resourceName	Świgra
resourceShortName	Świgra
downloadLocation	–
dateCreation	2003-2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN

contact.Person.surname	Woliński
contact.Person.givenName	Marcin
contact.Person.email	marcin.wolinski@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	to be defined
resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	nlpApplications
actualUse	nlpApplications
description	Świgra is a Prolog parser implementing Świdziński's Formal Grammar of Polish. Świgra uses a bottom-up parsing strategy, which for Polish proved to be superior to the top-down strategy. The parser builds a shared parse forest, which is not only the result but also a means of avoiding unnecessary recomputation. The rules of the grammar are not interpreted at the runtime but they are compiled to Prolog clauses.
relevantPublications	not yet available
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	pl
size	yet unknown
sizeUnit	–

resourceName	Anotatornia
resourceShortName	Anotatornia
downloadLocation	http://zil.ipipan.waw.pl/Anotatornia?action=AttachFile&do=view&target=anotatornia-2012-04-10-1206.tgz
dateCreation	2008-2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Lenart
contact.Person.givenName	Michał
contact.Person.email	michal.lenart@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	GPL v. 3
resourceLocation	http://zil.ipipan.waw.pl/Anotatornia/
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	npApplications
actualUse	npApplications
description	Anotatornia is a tool for the manual on-line annotation of corpora at various linguistic levels. The levels currently implemented are: word-level and sentence-level segmentation, morphosyntax, word sense disambiguation. Anotatornia implements sophisticated mechanisms of the management of texts, annotators and conflicts.

relevantPublications	Przepiórkowski A., Murzynowski G. Manual annotation of the National Corpus of Polish with Anotatornia. In: Stanisław Goźdź-Roszkowski, ed., Explorations across Languages and Corpora: PALC 2009, pp. 95-103, Frankfurt am Main, 2011. Peter Lang. Hajnicz E., Murzynowski G., Woliński M. ANOTATORNIA – lingwistyczna baza danych. In: Proceedings of the 5th conference InfoBazy 2008, Systems – Applications – Services, pp. 168–173, Sopot, 2008. Centrum Informatyczne TASK, Politechnika Gdańska.
resourceType	tool
mediaType	text
lingualityType	monolingual
languageId	pl
size	–
sizeUnit	–

resourceName	Ruler
resourceShortName	Ruler
downloadLocation	–
dateCreation	2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Ogrodniczuk
contact.Person.givenName	Maciej
contact.Person.email	maciej.ogrodniczuk@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	to be defined
resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	NlpApplications
actualUse	NlpApplications
description	Ruler is a rule-based coreference resolver for Polish. The implemented module uses standard best-first entity-based model based on syntactic constraints (elimination of nested nominal groups), syntactic filters (elimination of syntactic incompatible heads), semantic filters (wordnet-derived compatibility) and selection (weighted scoring). Syntactic properties are obtained from Spejd and its morphological component Morfeusz SGJP which produce NP chunks with detailed morphosyntatic information. Semantic properties are currently based on plWordNet.
relevantPublications	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.
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	pl
size	–

sizeUnit	–

resourceName	PolSumm
resourceShortName	PolSumm
downloadLocation	–
dateCreation	2003-2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Kulików
contact.Person.givenName	Sławomir
contact.Person.email	Slawomir.Kulikow@polsl.pl
DistributionInfo	available, unrestricted use
license	to be defined
resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	NlpApplications
actualUse	NlpApplications
description	PolSumm is a Polish document summarizer combining elements of a linguistic transformation of the text with statistical methods and information retrieval.
relevantPublications	not yet available
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	pl
size	yet unknown
sizeUnit	–

resourceName	VOICE LAB Automated Speech Recognition (ASR) engine
resourceShortName	VLASR
downloadLocation	not applicable
dateCreation	2011-2012
projectPartner	University of Łódź
iprHolder.organizationName	Voice Lab sp. z o.o.
contact.Person.surname	Szwelnik
contact.Person.givenName	Tomasz
contact.Person.email	tomasz.szwelnik@voicelab.pl
DistributionInfo	notAvailable
license	prioprietary
resourceLocation	http://www.voicelab.pl/

distributionAccessMedium	other
restrictionsOfUse	academic-nonCommercialUse commercialUse
licenseSignatory.Person.position	President
foreseenUse	NlpApplications
actualUse	NlpApplications
description	The VOICE LAB Automated Speech Recognition (ASR) engine enables recognition of natural speech. The ASR supports an industry standard known as Speech Recognition Grammar Specification (SRGS). The engine has been optimized for use in navigation of information kiosks, mobile applications, switch-boards or call centers supporting human operators. It is also used in voice search. The ASR can be used in as a service or as a standalone, on-site installation. The acoustic models oft he engine have been optimized for Polish. However, with an appropriate training sets, it can be used for any language as the core technology is language independent. The engine works on every Linux distribution, preferably a 64 bit one.
relevantPublications	not applicable
resourceType	technologyToolService
mediaType	audio
lingualityType	monolingual
languageId	pl
size	not applicable
sizeUnit	not applicable

resourceName	Language Detector
resourceShortName	LDetect
downloadLocation	http://pelcra.pl
dateCreation	2012
projectPartner	University of Łódź
iprHolder.organizationName	University of Łódź
contact.Person.surname	Piotr
contact.Person.givenName	Pęzik
contact.Person.email	piotr.pezik@uni.lodz.pl
DistributionInfo	
license	GPL
resourceLocation	–
distributionAccessMedium	downloadable
restrictionsOfUse	
foreseenUse	nlpApplications
actualUse	nlpApplications

description	The PELCRA language detector is a Java tool for detecting the language of an arbitrary stretch of text. The tool was developed by the PELCRA Team at the University of Łódź and it's available under the GPL licence. The first version 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.
relevantPublications	
resourceType	technologyToolService
mediaType	
lingualityType	multilingual
languageId	
size	
sizeUnit	

5.5. Serbian language resources detailed specification

resourceName	Media Multimedia Archive Ebart
resourceShortName	EbartMultimediaArchive
downloadLocation	
dateCreation	2003-
projectPartner	Ebart
iprHolder.organizationName	Ebart – Belgrade
contact.Person.surname	Ćurguz
contact.Person.givenName	Kazimir
contact.Person.email	office@archive.rs
DistributionInfo	underNegotiation
license	-
resourceLocation	http://www.arhiv.rs/
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	-
licenseSignatory.Person.position	-
foreseenUse	nlpApplications
actualUse	human use
description	The EbartMultimediaArchive database is a video archive that contains several hundred thousand broadcasts from the most important central TV stations and some local TV stations published since 2005. They are grouped using various critera (thematic, persons, etc.). A large number of them are transcribed to text.
relevantPublications	-
resourceType	corpus
mediaType	text video
lingualityType	monolingual
languageId	sr
size	500,000
sizeUnit	article

resourceName	Named Entities Evaluation Corpus for Serbian
resourceShortName	SrpNE-evaluation
downloadLocation	-
dateCreation	2010-
projectPartner	University of Belgrade, Faculty of Mathematics
iprHolder.organizationName	University of Belgrade, Faculty of Mathematics
contact.Person.surname	Krstev

contact.Person.givenName	Cvetana
contact.Person.email	cvetana@matf.bg.ac.rs
DistributionInfo	avaiable-restricted use
license	-
resourceLocation	-
distributionAccessMedium	Downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	-
foreseenUse	nlpApplications
actualUse	nlpApplications
description	This corpus consists of app. 3,000 short news in which named entities were automatically tagged and manually checked. NEs tagged are: persons, person roles and functions, temporal, measure and money expressions, organizations.
relevantPublications	Cvetana Krstev, Duško Vitas, Ivan Obradović, Miloš Utvić, "E-Dictionaries and Finite-State Automata for the Recognition of Named Entities", in Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing, FSMNLP 2011, Blois, France, July 12-15, 2010. eds. Andreas Maletti and Matthieu Constant, Association for Computational Linguistics, ISBN 978-3-642-14769-2, pp. 48-56, 2011.
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	sr
size	150,000
sizeUnit	word

resourceName	Semantically Tgged Corpus of Contemporary Serbian (preliminary version)
resourceShortName	-
downloadLocation	-
dateCreation	2012-
projectPartner	University of Belgrade, Faculty of Mathematics
iprHolder.organizationName	University of Belgrade, Faculty of Mathematics
contact.Person.surname	Vitas
contact.Person.givenName	Duško
contact.Person.email	vitas@matf.bg.ac.rs
DistributionInfo	avaiable-restricted use
license	-
resourceLocation	-
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse

licenseSignatory.Person.position	-
foreseenUse	nlpApplications
actualUse	nlpApplications
description	This corpus was semantically tagged on the basis of some semantic attributes associated to lemmas in Serbian e-dictionaries, as well as on Serbian Wordnet.
relevantPublications	-
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	sr
size	10,000
sizeUnit	word

resourceName	Serbian-English Aligned Literary Corpus
resourceShortName	-
downloadLocation	-
dateCreation	-
projectPartner	University of Novi Sad, Faculty of Philosophy
iprHolder.organizationName	
contact.Person.surname	Major
contact.Person.givenName	Randy
contact.Person.email	
DistributionInfo	underNegotiation
license	-
resourceLocation	-
distributionAccessMedium	-
restrictionsOfUse	-
licenseSignatory.Person.position	-
foreseenUse	npApplications
actualUse	human use
description	This aligned corpus consists of Serbian literary texts translated to English.
relevantPublications	-
resourceType	corpus
mediaType	text
lingualityType	Bilingual
languageId	en, sr
size	-
sizeUnit	word

resourceName	Terminological Database for Geology
resourceShortName	GeolISSTerm
downloadLocation	-
dateCreation	2006-
projectPartner	University of Belgrade, Faculty of Geology and Mining
iprHolder.organizationName	Ministry of Education and Sciences
contact.Person.surname	Stanković
contact.Person.givenName	Ranka
contact.Person.email	ranka@grf.bg.ac.rs
DistributionInfo	underNegotiation
license	-
resourceLocation	http://www.rgf.bg.ac.rs/
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	-
licenseSignatory.Person.position	-
foreseenUse	nlpApplications
actualUse	human use
description	The electronic dictionary of geologic terms (GeolISSTerm) is a special-purpose taxonomy of basic geologic concepts and terms. GeolISSTerm is an elementary electronic resource in the process of domain formation in the Geologic Information System of Serbia (GeolISS). It is the core of GeolISS through which validation, classification and specification of attributes of the observed and the interpreted takes place.
relevantPublications	Stanković, Ranka, and Branislav Trivić, and Olivera Kitanović, and Branislav Blagojević, and Velizar Nikolić. "The Development of the GeolISSTerm Terminological Dictionary." *INFOtheca* 12, 1: (2011) 49a-63a.
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	blingual
languageId	en, sr
size	3,500
sizeUnit	concept

resourceName	Emotion Classification of Serbian Texts
resourceShortName	
downloadLocation	-
dateCreation	2011-
projectPartner	University of Belgrade, Faculty of Mathematics
iprHolder.organizationName	University of Belgrade, Faculty of Mathematics
contact.Person.surname	Mladenović

contact.Person.givenName	Miljana
contact.Person.email	ml.miljana@gmail.com
DistributionInfo	underNegotiation
license	-
resourceLocation	http://cvetana.mmiljana.com
distributionAccessMedium	webExecutable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	-
foreseenUse	nlpApplications
actualUse	nlpApplications
description	This system for emotion classification of Serbian texts is based on an ontology built specially for this purpose that functions as an emotion classifier. It is based on well-known discrete emotions theories of Arnold, Ekman, Frijda, Gray, Izard, Tomkins, Weiner&Graham, Watson and Plutchik. Each of these theories reviews human emotions as discrete and independent and describes them by small bag of words. These words are used to build the emotions ontology. In order to expand the extraction of information from texts a Serbian associative-dictionary was used coupled with Serbian morphological electronic dictionaries yielding some nine thousand forms used by the system. Extracted RDF structures are then submitted for reasoning and frequencies of emotions are calculated according to each of theories individually. Finally, for the visual presentation of results a separate graphical unit was created. The application is realized on Csharp Net Framework platform.It can be tested on texts in .html and .txt formats and it accepts both Cyrillic and Latin scripts. Text files can be manually pasted, uploaded from a local system or used directly from a given URL address on Web.
relevantPublications	-
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	sr
size	-
sizeUnit	-

resourceName	Named Entities Module for Serbian
resourceShortName	SrpNE-module
downloadLocation	-
dateCreation	2009-
projectPartner	University of Belgrade, Faculty of Mathematics
iprHolder.organizationName	University of Belgrade, Faculty of Mathematics
contact.Person.surname	Krstev
contact.Person.givenName	Cvetana
contact.Person.email	cvetana@matf.bg.ac.rs

DistributionInfo	underNegotiation
license	-
resourceLocation	-
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	-
foreseenUse	human use nlpApplications
actualUse	human use nlpApplications
description	This module for named entity recognition and tagging is based on Serbian morphological e-dictionaries and a large collection of Finite-State Transducers (in the form of cascades). It recognizes and tags: persons, person roles and functions, temporal expressions, mount expressions (including measures and money expressions) and organizations. The module is integrated in a web service and tags NEs in texts uploaded by users.
relevantPublications	Cvetana Krstev, Duško Vitas, Ivan Obradović, Miloš Utvić, "E-Dictionaries and Finite-State Automata for the Recognition of Named Entities", in Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing, FSMNLP 2011, Blois, France, July 12-15, 2010. eds. Andreas Maletti and Matthieu Constant, Association for Computational Linguistics, ISBN 978-3-642-14769-2, pp. 48-56, 2011.
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	sr
size	-
sizeUnit	

resourceName	A web tool for aligned text search
resourceShortName	
downloadLocation	-
dateCreation	2012-
projectPartner	University of Belgrade, Faculty of Mathematics
iprHolder.organizationName	University of Belgrade, Faculty of Mathematics
contact.Person.surname	Zečević
contact.Person.givenName	Anđelka
contact.Person.email	andjelkaz@matf.bg.ac.rs
DistributionInfo	underNegotiation
license	-
resourceLocation	-
distributionAccessMedium	accessibleThroughInterface

restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	
foreseenUse	human use
actualUse	human use
description	This is a web tool for effective search of aligned and annotated texts. It is especially designed for texts in which named entities were tagged. Its purpose is to compare annotation of NEs in aligned text and for that purpose a language independent classification schema for NEs is used.
relevantPublications	-
resourceType	technologyToolService
mediaType	text
lingualityType	multilingual
languageId	-
size	-
sizeUnit	-

resourceName	Web Applications (NE extraction from web pages)
resourceShortName	
downloadLocation	-
dateCreation	2012-
projectPartner	University of Belgrade, Faculty of Mathematics
iprHolder.organizationName	University of Belgrade, Faculty of Mathematics
contact.Person.surname	Vitas
contact.Person.givenName	Duško
contact.Person.email	vitas@matf.bg.ac.rs
DistributionInfo	underNegotiation
license	-
resourceLocation	-
distributionAccessMedium	other -executable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	-
foreseenUse	nlpApplications
actualUse	nlpApplications
description	This is a web tool for extraction of proper names from categories given in Wikipedia for English, French, Serbian, Polish.
relevantPublications	-
resourceType	technologyToolService
mediaType	text
lingualityType	multilingual
languageId	en, fr, sr, pl
size	-
sizeUnit	-

resourceName	Language Model for Serbian
resourceShortName	-
downloadLocation	-
dateCreation	2012
projectPartner	Ebart - Belgrade
iprHolder.organizationName	Ebart - Belgrade
contact.Person.surname	Ćurguz
contact.Person.givenName	Kazimir
contact.Person.email	office@archive.rs
DistributionInfo	underNegotiation
license	-
resourceLocation	http://www.arhiv.rs/
distributionAccessMedium	downloadable
restrictionsOfUse	commercialUse
licenseSignatory.Person.position	-
foreseenUse	nlpApplications
actualUse	nlpApplications
description	This language model of Serbian is produced on the basis of the large newspaper corpus (approx. 4 million articles) using the standard methodology for such models.
relevantPublications	-
resourceType	language description
mediaType	text
lingualityType	monolingual
languageId	sr
size	-
sizeUnit	-

5.6. Slovak language resources detailed specification

resourceName	Database of Root Morphemes
resourceShortName	Database of root morphemes
downloadLocation	
dateCreation	2012
projectPartner	Prešov University
iprHolder.organizationName	Prešov University
contact.Person.surname	-
contact.Person.givenName	-
contact.Person.email	-
DistributionInfo	underNegotiation
license	
resourceLocation	Prešov University
distributionAccessMedium	-
restrictionsOfUse	-
licenseSignatory.Person.position	-
foreseenUse	nlpApplications
actualUse	nlpApplications
description	Database provides alternative approach to morphology analysis. It contains 67,000 linguistic units with deep morphematic linguistic analysis. It has been compiled at the Prešov University in Prešov and has been used as a basis for a published Slovník koreňových morfém slovenčiny.
relevantPublications	Slovník koreňových morfém slovenčiny. M. Sokolová et al. ISBN 9788080683191.
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	monolingual
languageId	sk
size	67,000
sizeUnit	root morpheme

resourceName	Dictionary of Slovak Adjective Collocations
resourceShortName	Dictionary of Slovak Adjective Collocations
downloadLocation	
dateCreation	2012
projectPartner	LSIL
iprHolder.organizationName	University of St. Cyril and Methodius in Trnava
contact.Person.surname	-
contact.Person.givenName	-
contact.Person.email	-
DistributionInfo	underNegotiation

license	-
resourceLocation	University of St. Cyril and Methodius in Trnava
distributionAccessMedium	
restrictionsOfUse	-
licenseSignatory.Person.position	
foreseenUse	nlpApplications
actualUse	nlpApplications
description	The dictionary provides an overview of the combinatorial behaviour of words and contains collocation profiles of the most frequent Slovak adjectives. The combinatorial potentials of word forms of a word are the basis for the creation of so-called collocational templates which the patterns of collocations are based on. The dictionary is currently being compiled (currently, it contains collocation profiles of 140 adjectives). The dictionary is being created at the University of St. Cyril and Methodius in Trnava, with input from the Ľ. Štúr Institute of Linguistics.
relevantPublications	
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	monolingual
languageId	sk
size	140
sizeUnit	entry

resourceName	Dictionary of German-Slovak Collocations
resourceShortName	Dictionary of German-Slovak Collocations
downloadLocation	-
dateCreation	2012
projectPartner	
iprHolder.organizationName	University of St. Cyril and Methodius in Trnava
contact.Person.surname	-
contact.Person.givenName	-
contact.Person.email	-
DistributionInfo	underNegotiation
license	
resourceLocation	University of St. Cyril and Methodius in Trnava
distributionAccessMedium	-
restrictionsOfUse	-
licenseSignatory.Person.position	
foreseenUse	nlpApplications
actualUse	nlpApplications
description	Dictionary of German-Slovak Collocations provides confrontational overview of the combinatorial behaviour of words in bilingual comparison. The database consists of German collocations (currently 440 profiles) with Slovak equivalents The dictionary is being created at the University of St. Cyril and Methodius in Trnava.

relevantPublications	
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	bilingual
languageId	DE, SK
size	440
sizeUnit	entry

resourceName	Multimodal Multilingual Dictionary of Gestures
resourceShortName	DiGest
downloadLocation	-
dateCreation	2012
projectPartner	Institute of Informatics, Slovak Academy of Sciences
iprHolder.organizationName	Institute of Informatics, Slovak Academy of Sciences
contact.Person.surname	-
contact.Person.givenName	-
contact.Person.email	-
DistributionInfo	underNegotiation
license	
resourceLocation	Institute of Informatics, Slovak Academy of Sciences.
distributionAccessMedium	-
restrictionsOfUse	-
licenseSignatory.Person.position	
foreseenUse	nlpApplications
actualUse	nlpApplications
description	DiGest contains a database of extra-verbal expressions. Its current version contains several hundreds of gestures represented by a still image, a description of the gesture and its meaning, and optional sound and video records. The current version includes language and culture dependent content for American English, Slovak, Italian, and Mongolian. Entries for Japanese, Chinese, and Hungarian are also included. The database has been compiled at the Institute of Informatics, Slovak Academy of Sciences.
relevantPublications	
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	multilingual
languageId	en, it, jp, cn, hu
size	324
sizeUnit	entry

resourceName	Language model prim-5.0-inf
resourceShortName	

downloadLocation	http://korpus.sk/prim(2d)5(2e)0(2f)models.html
dateCreation	2012-02-01
projectPartner	LSIL
iprHolder.organizationName	LSIL
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available-unrestricted use
license	the Open Database License v1.0
resourceLocation	LSIL
distributionAccessMedium	downloadable
restrictionsOfUse	ShareAlike, attribution
licenseSignatory.Person.position	Director
foreseenUse	nlpApplications
actualUse	nlpApplications
description	This is a language model of journalistic style. The model is in iARPA format, using witten-bell smoothing. It was created by the IRSTLM Tooklit. The model is lowercased. It has been released with the contribution of the EuroMatrixPlus project.
relevantPublications	-
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	sk
size	515,000,000
sizeUnit	token

resourceName	Language model prim-5.0-vyv
resourceShortName	
downloadLocation	http://korpus.sk/prim(2d)5(2e)0(2f)models.html
dateCreation	2012-02-01
projectPartner	LSIL
iprHolder.organizationName	LSIL
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available-unrestricted use
license	the Open Database License v1.0
resourceLocation	LSIL
distributionAccessMedium	downloadable
restrictionsOfUse	ShareAlike, attribution

licenseSignatory.Person.position	director
foreseenUse	nlpApplications
actualUse	nlpApplications
description	A language model from the Slovak National Corpus. The model is in iARPA format, using witten-bell smoothing. It was created by the IRSTLM Tooklit. The model is lower-cased. It has been released with the contribution of the EuroMatrixPlus project.
relevantPublications	-
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	sk
size	247,000,000
sizeUnit	token

resourceName	Language model prim-5.0-sane
resourceShortName	
downloadLocation	http://korpus.sk/prim(2d)5(2e)0(2f)models.html
dateCreation	2012-02-01
projectPartner	LSIL
iprHolder.organizationName	LSIL
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available-unrestricted use
license	the Open Database License v1.0
resourceLocation	LSIL
distributionAccessMedium	downloadable
restrictionsOfUse	ShareAlike, attribution
licenseSignatory.Person.position	director
foreseenUse	nlpApplications
actualUse	nlpApplications
description	A language model from the Slovak National Corpus. The model is in iARPA format, using witten-bell smoothing. It was created by the IRSTLM Tooklit. The model is lowercased. It has been released with the contribution of the EuroMatrixPlus project.

relevantPublications	-
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	sk
size	733,000,000
sizeUnit	token