CESAR

Central and South-East European Resources
Project no. 271022

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

Version No. 1.4
03/11/2011

Document Information

Deliverable number:	D2.3
Deliverable title:	Report on resources (actually or potentially) available to the consortium (name)
Due date of deliverable:	29/07/2011
Actual submission date of deliverable:	03/11/2011
Main Author(s):	Svetla Koeva (IBL)
Participants:	Tamas Varadi (HASRIL) Tibor Pinter (HASRIL) Radovam Garabik (LSIL) Maciej Ogrodniczuk (IPIPAN) Adam Przepiórkowski (IPIPAN) Marko Tadic (FFZG) Dusko Vitas (UBG)
Internal reviewer:	Tamas Varadi (HASRIL)
Workpackage:	2
Workpackage title:	Analysis and selection of language resources
Workpackage leader:	IBL
Dissemination Level:	Public
Version:	1.4
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
1.4	03/11/2011	modificated	Svetla Koeva		Modification upon the request of IPIPAN
1.3	31/10/2011	Final	Tamás Váradi	supervision	update after adding new tasks to the project by IPIPAN
1.2	29/07/2011	draft	Tibor Pinter	editing of the text	
1.1	29/07/2011	draft	Svetla Koeva	Drafting report	

EXECUTIVE SUMMARY

The deliverable gives a detailed description on the actually or potentially available resources to the consortium. 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 first six month of the project.

Table of Contents

1. Background

1.1. Project objectives

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

The main goals of 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, inter-connected repositories, exchanges, facilities etc. that can be used by language researchers, developers and professionals;
- to mobilise 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 the other parts of Europe by filling obvious and important gaps in language resources and tools infrastructure.

1.2. Baseline situation

The CESAR project will specifically focus 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) being official languages of recently joined member states, while two languages (Croatian and Serbian) represent 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 from these languages. Building on a wide range of already existing resources and previous national or international activities, the project will create, populate and operate a comprehensive language-resource platform enabling and supporting large-scale multi- and cross-lingual products and services. In extensive cooperation with META-NET, resources will be upgraded and updated to widely acknowledged standards, thus ensuring interoperability and creating the ground for widespread and efficient and the potential to modularize them in language technology pipelines.

In the frame of this task language resources and tools already developed or still under development have been and will be identified. The D2.3 Report on resources (actually or potentially) available to the consortium represents the resources for Bulgarian, Croatian, Hungarian, Polsih, Serbian and Slovak identified so far.

1.3. Target resources and users

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

The target users are developers and researchers both in industry and academia. This includes 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, EC institutions, and private investors.

2. A common and shared resource description

CESAR supports the goal of a common and shared resource description between the four projects constituting METANET (i.e. CESAR, METANET4U and META-NORD, and T4ME). The focus was to gather all relevant information (metadata) of the resources actually (or potentially) available. This metadata covers features of the localization of the resources, information on IPR holders (the name of the holder as well as the addresses of the main contact person), the distribution of the media (the specified the format used for the delivery of the resource), as well as the licence issues and restrictions of its usage. The metadata also describes the NLP focused usage of the resources both in its actual and in its upcoming state (actual and foreseen usage). The metadata contains wider information of the resources by offering further readings and publications on the resources, as well as links of their main documentation. The metadata scheme of the resources 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 in T4ME/META-NET

CESAR adopted the metadata scheme developed in T4ME/META-NET - thereby a common metadata description for language resources in many different European languages will be provided. The Table 1 below describes the metadata scheme with definitions and recommended values used in T4Me and shared by other four projects part of META-NET.

	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 more than one contact persons 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	Specifies 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; NLP applications

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; NLP applications

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
multilinguality Type	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 values 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. META-NET metadata scheme

2.2. Project specific additions to the scheme

In addition in the CESAR project some new metadata fields are accepted for the metadata scheme. They are as follows - 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

3. Resources identified via CESAR by month sixth

The D2.3 Report on resources (actually or potentially) available to the consortium gives an overview of the main language resources of the Central-East Europe. It is compiled to give more than 30 types of information on resources of six languages. A table containing values of the commonly accepted metadata scheme was constructed by a survey on national level with help of 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, and offers a catalogue of written and spoken language resources that will be contributed to the project.

The description gives a detailed view of the main language resources available on languages covered by the partners of the project. The description contains language resources on 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:

3.1.1. Text corpora

Monolingual corpora

BulNC – The Bulgarian National Corpus is a large-scale, representative, publicly available corpus of Bulgarian. It is a monolingual general corpus, fully morpho-syntactically (and partially semantically) annotated. Presently, the BulNC consists of about 450 000 000 tokens and includes more than 10 000 samples.

BulPoSCor - The POS Tagged Bulgarian Corpus is the result of the manual POS disambiguation of each wordform by language experts. The POS Tagged Bulgarian Corpus comprises more than 150 thousand words. It was designed through the extraction of samples of 300 words minimum (expanded to sentence boundary) from the Structured Brown Corpus of Bulgarian. Parts of the POS Tagged Corpus were used as training and test corpora in the creation of several Bulgarian taggers. The POS Tagged Bulgarian Corpus enables efficient search of language patterns and forms in the text.

BulSemCor - The Sense Tagged Bulgarian corpus contains sense-disambiguated lexical items defined in the context of occurrence. It consists of excerpts from the samples from the "Brown" Corpus of Bulgarian and preserves its overall structure: excerpts of minimum 100 words were clipped according to a methodology accounting for the concentration of the most frequent words from different classes. As a result the input corpus for semantic annotation amounts to app. 100 thousand words since the text excerpts were expanded to the left and right to sentence boundaries. Each lexical item (simple or compound word) in BulSemCor is assigned manually the unique semantic or grammatical meaning from the Bulgarian wordnet

which occurs in the particular context. BulSemCor is used as a training and test set in the elaboration of a probability formalism for automatic word-sense disambiguation oriented towards machine translation.
Bilingual and Multilingual (with Bulgarian as one language)

Bul-XCor – The Bulgarian-X language parallel corpora contains texts that are Bulgarian translations of English originals and translations from a third language into both English and Bulgarian. The most important condition for a text to be selected for the corpus was that it was either created after 1945 or its Bulgarian translation was made after that year. The fiction subcorpus consists of 340 samples texts per language – 34 553 474 words for the Bulgarian part and 39 590 472 words for the English part. All together the Bulgarian samples approximate 100 000 000 words.

3.1.2. Lexical Conceptual Resources:

BulNet – The Bulgarian WordNet represents an electronic lexical semantic network, containing synsets with glosses and 17 more relations, for the most part semantic ones, such as antonymy, meronymy, causation, category domain, etc. Currently the Bulgarian database amounts to 35 000 synsets and 54 700 conceptual, morpho-semantic and language external relations. Beside intralingual ones wordnet also encodes interlingual relations which associate the equivalent synsets in 18 European languages. It is through these relations that BulNet becomes part of the Global WordNet - a conceptual ontology of concepts lexicalized in natural languages and their relations. Wordnet has been successfully used in intelligent information search and information retrieval from documents in different languages, text categorisation and text summarisation, word sense disambiguation, machine translation, as well as in many other NLP tasks.

BulFrameNet - The Bulgarian FrameNet represents general semantic and language-specific lexical-semantic and syntactic combinatory properties of Bulgarian lexical units. BulFrameNet database so far contains unique descriptions of over 3 000 Bulgarian lexical units, approx. one tenth of them aligned with appropriate semantic frames. Each lexical entry consists of a lexical unit; a semantic frame from the English FrameNet, expressing abstract semantic structure; a grammatical class, defining the inflexional paradigm; a valency frame describing (some of) the syntactic and lexical-semantic combinatory properties (an optional component); and (semantically and syntactically) annotated examples.

BulGram - The Grammar Dictionary of Bulgarian is an inflexion dictionary that consists of approximately 85 000 lemmas and allows automatic word form analysis and generation resulting in approximately 1 mln. 140 thousand word forms. The formal structure of the dictionary is based on finite state transducers (FSTs) that are widely applied in up-to-date electronic dictionaries. In its present state BulGram represents in full the synthetic word formation patterns in Bulgarian and is therefore suitably transformed into spelling checking dictionary.

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	annotationType
Bulgarian National Corpus	BulNC	http://search.dcl.bas.bg	corpus	450 000 000	token	MSD, sense
Bulgarian PoS Annotated Corpus	BulPoSCor	http://search.dcl.bas.bg	corpus	150 000	word	MSD

Bulgarian Sense Annotated Corpus	BulSemCor	http://search.dcl.bas.bg	corpus	105 000	word	sense
Bulgarian-X language parallel corpora	Bul-XCor	http://search.dcl.bas.bg	corpus	100 000 000	token	MSD, sentence alignment
Bulgarian wordnet	BulNet	http://catalog.elra.info/product_info.php?products_id=802	lexicalConceptualResource	35 000	synset	PWN
Bulgarian FrameNet	BulFrameNet	http://dcl.bas.bg/LexIt	lexicalConceptualResource	3 000	lemma	FN
Bulgarian Grammatical Dictionary	BulGram	http://dcl.bas.bg/est	lexicalConceptualResource	85 000	lemma	

Table 3.2 Summary of Croatian language resources

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 still in development or are constantly updated, can be classified in the following categories:

3.3.1. Text corpora

Monolingual (Hungarian) corpora

HNC - The Hungarian National Corpus a the general purpose, representative corpus of today's written Hungarian language. It gives an exact, quantifiable picture of Hungarian language use. It includes bibliographical metadata, and encodes the boundaries of structural units (paragraphs, sentences). The corpus is automatically POS-tagged: we have lemma, part of speech, and morphological analysis for each word. The query interface is freely available to anyone. It incorporates Hungarian texts from Hungary, Slovakia, Transcarpathia, Transylvania and Vojvodina in five different genres: press, literature (from Digital Literature Academy), scientific, official and personal communication. The size of the corpus is currently 187 million words. The partially syncactically parsed form of the whole corpus can be queried by the "Verb Argument Browser" tool: http://corpus.nytud.hu/vab

Szeged NER corpus - The Szeged NER corpusis a manually annotated part of the Szeged treebank, consisting of short business news. The used NER categories are (based on the CoNLL system (http://www.cnts.ua.ac.be/conll2003/ner/)) the following: PERSON, ORGANISATION, LOCATION and OTHER.

Szeged corpus - morpho-syntactically annotated and manually disambiguated corpus of 1,2 million words.

Hungarian webcorpus - over 1.48 billion words unfiltered (589m words fully filtered), this is by far the largest Hungarian language corpus, and it is available in its entirety under a permissive Open Content license. The Hungarian webcorpus was created as part of the WordSword project at the Media Research and Education Centre. The Webcorpus may be downloaded in two formats: as a frequency dictionary based on the texts and as the original texts.

Bilingual and Multilingual (with Hungarian as one language)

Hunglish - The Hunglish Parallel Corpus is a free sentence-aligned Hungarian-English parallel corpus of about 2 million sentences. The corpus may be searched through a web-based sentence search service. This service has more than 200,000 visits per month.

3.3.2. Lexical Conceptual Resources

Hungarian Wordnet - The Hungarian WordNetis a multilingual ontology, meaning that most of its synsets were mapped to equivalent concepts in English (Princeton) WordNet v. 2.0. The ontology is also linked to entries of a Hungarian Monolingual explanatory dictionary and to the entries of the Hungarian verb valency frame lexicon.

Szeged treebank - manually checked treebank of 1,2 million words.

3.3.3. Speech databases

BABEL - BABELHungarian Clear Speech Databaseis an EUROM1 compatible database containing records from 60 speakers distributed in 3 speakers set (many, few, very few).

MRBA – The Hungarian Reference Speech Database continuous read speech. During the planning of the corpus, we took into consideration the special characteristics of Hungarian language. Since the Hungarian is an agglutinative language, we needed to create a larger vocabulary in some categories, than it is mandatory. We tried to pay an extra attention to the topic 'phonetically rich sentences and words', to create a phonetically well balanced speech database for text independent speech recognizers. The database contains utterances read by 332 different speakers. The utterances were recorded in acoustically different locations.

MTBA – The Hungarian Telephone Speech Database is a PSTN and mobil telephone voice Hungarian speech database. The database contains records based on the definition in SpeechDatE for the dialectical, age and sex balance and vocabulary. Important and different from the SpeechDatE database is, that the phonetically rich sentences and words have been segmented and labelled at phoneme level. The database has two parts. The first part (2 CDs) contains labeled application words, numbers, dates, spelling and names, the second part (1 CD) contains labeled and segmentated phonetically rich sentences and words.

MTÜBA - Hungarian Telephone Client Speech contains telephone calls recorded at the call centre of a service provider company. The database will be fully anonimized. The corpus consists of dialogues between the operator and the client. The ortographic transcription of the speech utterances is provided, clauses are segmentated (automatically followed by hand-correction). Parts of speech holding emotions acoustically audible are also labelled according to 4 basic emotions.

Emotion Database - Spoken databse holding emotionally reach utterances, labelling is done for emotions (8 basic emotions are labelled)

Sound Gesture Database - Database of sound gestures consisting of 770 tokens.

Medical Database - The Medical databaseis a speech corpus holding utterances from persons suffering from different speech problems of organic origine.

Hungarian MALACH - Hungarian Speech Database of Holocaust Survovors' Testimonies histories of elderly people from the world war II, typically they are holocaust survivors. Originally videos were recorded with 2 microphones on two channel, but in Aitia International only 1 channel audio files are available with transcriptions.

Hungarian Parliamentary Speeches - publicly available corpus with approximate transcriptions of Parliamentary Speeches. In this project, time alignment will be made and not alignable part will be marked.

3.3.4. Multimedia databases

BND - Broadcast News Databasewas collected as a member of the Broadcast News Interest Group of COST278, the COST action on Speech and Language Interaction in Telecommunications in cooperation of 10 different institutions throughout Europe. The Hungarian material consists of 3h and 30 minutes of recordings, transcribed and annotated, using the conventions of NIST (National Institute of Standards and Technology, USA).

BLD or ME corpus (ME: Mindentudás Egyeteme) – TheBroadcast Lectures Database contains recorded Broadcast Video Lectures from wide scientific topics for the public.

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	annotationType
Hunglish parallel corpus	Hunglish	hosted by owner	corpus	54200000	token	segmentation; alignment
Hungarian Wordnet		hosted by the consortium	lexical conceptual resource	42000	synset	PWN
Hungarian National Corpus	HNC	hosted by owner	corpus	187600000	token	MSD
Szeged NER corpus		hosted by owner	corpus	220000	token	POS, NE
Szeged corpus		hosted by owner	corpus	1200000	token	POS (MSD)
Szeged treebank		hosted by owner	corpus	1200000	token	POS (MSD), treebank
Hungarian webcorpus		hosted by owner	corpus	1480000000	token	none
BABELHungarian Clear Speech Database	BABEL	hosted in international repository (ELRA)	corpus	4	hour	annotation, segmentation
Hungarian Reference Speech Database	MRBA	hosted by the consortium	corpus	6,5	hour	annotation, segmentation
Hungarian Telephone Speech Database	MTBA	hosted by the consortium	corpus	5	hour	annotation, segmentation
Hungarian Telephone Client Speech Database	MTÜBA	hosted by owner	corpus	60	hour	annotation
Broadcast News Database	BND	hosted by owner	corpus	5	hour	annotation
Emotion Database		hosted by owner	corpus	50	hour	annotation
Sound Gesture Database		hosted by owner	lexical conceptual resource	770	token	annotation

Medical Database		hosted by the consortium	corpus	1	hour	annotation
Broadcast Lectures Database	BLD or ME corpus	hosted by owner	corpus	~150+	hour	transcription
Hungarian Speech Database of Holocaust Survovors' Testimonies	Hungarian MALACH	AITIA International	corpus	31	hour	transcription
Hungarian Parliamentary Speeches			corpus	1000+	hour	alignment, annotation

Table 3.3 Summary of Hungarian language resources

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 still in development or are constantly updated, can be classified in the following categories:

3.4.1. Text corpora

Monolingual (Polish) corpora

NKJP - **National Corpus of Polish** a shared initiative of four institutions: Institute of Computer Science at the Polish Academy of Sciences (coordinator), Institute of Polish Language at the Polish Academy of Sciences, Polish Scientific Publishers PWN, and the Department of Computational and Corpus Linguistics at the University of Łódź. It has been registered as a research-development project of the Ministry of Science and Higher Education. The list of sources for the corpus contains classic literature, daily newspapers, specialist periodicals and journals, transcripts of conversations, and a variety of short-lived and internet texts. The resources represent wide diversity with respect to the subject and genre. The spoken part covers both male and female speakers, in various age groups, coming from various regions in Poland.

WKSF - The corpus of frequency dictionary of Polish language of the XX century sixties was collected for the purpose of creating general frequency dictionary of contemporary Polish. The work started in 1967. Partial results were published between 1972 and 1977, the completed dictionary in 1990. The corpus was later augmented in various respects, both by manual editing and automated procedures. Corpus data contain 10,000 samples divided into 5 parts: essays, news, scientific texts, fiction and plays. Every sample is approximately 50 words long, they all come from texts published between 1963 and 1967 and contain bibliographic description of its source. Each word is tagged with its base form and some morphological properties. Sentence boundaries are also marked. Currently the corpus undergoes manual verification of the morphological descriptions.

Bilingual and Multilingual (with Polish as one language)

PPC - Polish Parallel Corpora will include various parallel resources, including Acquis Communautaire, Europarl, OPUS (EMEA, KDE, Open Subtitles), Parasol, CORDIS and RAPID news etc.

3.4.2. Lexical Conceptual Resources

PoliMorf morphological dictionary - two most important morphological dictionaries of Polish – Morfeusz SGJP and Morfologik. Morfeusz was developed by Zygmunt Saloni (author of linguistic data used in the analyser) and Marcin Woliński (programming part). The result morphological codes come from the IPI PAN Tagset (currently de facto standard for Polish) developed by Marcin Woliński and Adam Przepiórkowski for the annotation of the IPI PAN Corpus of Polish. Morfologik is based on current ispell dictionaries and Java libraries interfacing them. The result tags come from IPI Tagset.

plNER - NE resources with gazetteers - The gazetters has been obtained from existing sources and supplemented with additional language-specific resources acquired from the Web. Whenever appropriate, inflected forms were generated using Morfeusz SGJP generator. Gazetteer data was used in the process of named entity annotation of NKJP.

plWordNet - Polish WordNet (Słowosieć) is network of lexical-semantic relations, an electronic thesaurus with a structure modelled on that of the Princeton WordNet and those constructed in the EuroWordNet project. Polish WordNet describes the meaning of a lexical unit of one or more words by placing this unit in a network of links which represent such relations as synonymy, hypernyny, meronymy etc. To reduce the cost of the project, Polish WordNet has been built semi-automatically. Lexical relations were automatically recognized in large corpora of Polish and suggested to linguists/lexicographers via a graphical interface.

Polish Treebank - contains trees created with Świgra – a deep parser of Polish implemented by Marcin Woliński on the basis of a metamorphosis grammar of Polish GFJP created by Świdziński.

Polish Valency Dictionary (PVD) – a new resource resulting from merger of several valency dictionaries that have been made available in 2009 (a national Ministry of Science and Higher Education project "Automatic detection of semantic dependencies within verb argument structures in large treebanks") and in 2011 (a national Ministry of Science and Higher Education research grant "Construction of a treebank for Polish using automatic syntactic analysis") with an older, but very popular dictionary by prof. Świdziński.

Cross-lingual Repository of Named Entities (CRON) – a new multilingual dictionary of named entities prepared with co-operation of the University of Tours.

3.4.3. Speech database

PSCC - Polish Spoken Conversational Corpus originally a sub-corpus of the spoken part of the PELCRA Reference Corpus of Polish, which has been further expanded withing the National Corpus of Polish project. The corpus is composed of spontaneous natural conversations. After obtaining the participants' permission, the recordings were carefully transcribed and annotated with information about the respondents' sex, age, and educational level. All the original recordings have been digitized so that the conversation may not only be read, but also heard. At present the corpus contains some 1 800,000 words of transcribed speech accompanied by 75 hours of digitized recordings.

3.4.4. Multimedia databases

PSC - The Polish Sejm Corpuswill contain utterances of Polish Sejm members from cadencies 1-6. It will be morphologically annotated and made available for search with NKJP search tool – Poliqarp.

SMC – The Polish Spoken Multimedia Corpus contains transcripts of spontaneous informal conversations with the original recording files. It is available as TEI P5 encoded spoken Polish data.

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	annotationType
Polish Sejm Corpus	PSC	http://clip.ipipan.waw.pl/PSC	corpus	100000000	words	POS, MSD, chunking, NE
PoliMorf morphological dictionary	PoliMorf	http://clip.ipipan.waw.pl/PoliMorf	lexicalConceptualResource	3000000	words	POS, MSD
Polish Treebank	Treebank	http://clip.ipipan.waw.pl/plTreebank	corpus; lexicalConceptualResource	6000	sentences	POS, MSD, chunking, deep parsing
1M National Corpus of Polish	1MNKJP	http://www.nkjp.pl	corpus	1M	words	POS, MSD, chunking, word sense, NE
Polish Parallel Corpora	PPC	http://clip.ipipan.waw.pl/plPPC	corpus	2000000	words	segmentation
Polish Spoken Multimedia Corpus	SMC	http://clip.ipipan.waw.pl/plSMC	corpus	2500000	words	POS
Polish Spoken Conversational Corpus	PSCC	http://www.nkjp.uni.lodz.pl/spoken.jsp	corpus	1500000	words	segmentation, discourse, sociolinguistic
The corpus of frequency dictionary of Polish language of the XX century sixties	WKSF	http://clip.ipipan.waw.pl/PL196x	corpus	500000	words	segmentation, POS, MSD
NE resources with gazetteers	plNER	http://clip.ipipan.waw.pl/plNER	lexicalConceptualResource	229681	word forms	-
Polish WordNet (Słowosieć)	plWordNet	http://plwordnet.pwr.wroc.pl/wordnet/	lexicalConceptualResource	20223	lemmas (17695 synsets)	-
Polish Valency Dictionary	PVD	http://clip.ipipan.waw.pl/PVD	lexicalConceptualResource	1500	entries	valency information
Cross-lingual Repository of Named Entities	CRON	http://clip.ipipan.waw.pl/CRON	lexicalConceptualResource	10000	entries	LMF

Table 3.4 Summary of Polish language resources

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 still in development or are constantly updated, can be classified in the following categories:

3.5.1. Text corpora

Monolingual (Serbian) corpora

SrpKor - Corpus of Contemporary Serbian – developed by the NLP group at the Faculty of Mathematics, University of Belgrade (MATF-UBG) was partly funded by the Ministry of Education and Science. The list of sources for the corpus contains literature, monographs, daily newspapers, specialist periodicals and journals, manuals and textbooks. The resources represent wide diversity with respect to the subject and genre. It can be accessed via web interface. Its size is 113 million words.

SrpLemKor - Serbian MSD Annotated Corpus – is being developed by the same group. It will consist of a sample of various texts from SrpKor. It will be lemmatized and MSD tagged. It will contain 5 million words and it will be accessible via web interface at November 2011.

AlfaNumKor - AlfaNum Text Corpus of Serbian – was developed by the research group specialized for speech technologies that works at the Faculty of Technical Sciences and in "AlfaNum - Speech Technologies" in Novi Sad. This resource is developed for the purpose of linguistic research and development of commercial applications of speech and other language technologies. It consists of 100,517 lexemes (3,888,407 inflected forms).

CSL - Corpus of Serbian Language – was developed at the Faculty of Philosophy, University of Belgrade. It was compiled from a sample of 11 million words and spans the Serbian language from the 12th century to the present day. Each word in the CSL is manually tagged for its grammatical status (at the level of inflected morphology), number of graphemes and syllables and phonological structure. The system of tagging consists of about 2000 grammatical (inflected) forms.

Bilingual and Multilingual (with Serbian as one language)

SrpFranKor - French-Serbian Aligned Corpus - includes French or Serbian source literary texts and their translations. Texts are segment aligned and manually checked to obtain one-to-one alignment. Its size is approximately one million and a half words in each language.

SrpEngKor - English-Serbian Aligned Corpus - includes English or Serbian source texts from the domains: education, health, legislation, jurisprudence. Texts are segment and lemmatized and PoS tagged, aligned and manually checked. It will be accessible via web interface. It contains one million words per language.

Verne80days - Multilingual Edition of Verne's Novel "Around the World in 80 Days" - contains 25 translations of Jules Verne's novel "Around the World in 80 Days". Presently 15 of these translations are one-to-one segment aligned with the source text, and the alignment is manually checked. All other translations will also be aligned.

3.5.2. Lexical Conceptual Resources

SrpWN - Serbian Wordnet – developed by the NLP group at MATF-UBG represents a lexical semantic network, containing synsets with glosses and various semantic relations, such as antonymy, meronymy, causation, category domain, etc. Currently the Serbian Wordnet contains near to 16,000 synsets. The initial version of the Serbian Wordnet was produced in the scope of the EU-funded Balkanet project. Through interlingual relations it is connected to English Wordnet, and wordnets of many other languages. It will be accessible via web interface.

SrpRec - Serbian Morphological Dictionary – developed by the same group is a dictionary that consists of approximately 90 000 simple lemmas and 6,500 multi-word lemmas that allows automatic generation of inflected forms using finite state transducers (approximately 4 million simple word forms). All produced word forms are supplied by detailed morpho-syntactic description. Besides that, all lemmas are equiped with semantic, syntactic, domain and other codes.

SrpNER - Serbian Named Entity Resources – developed by the same group comprise of dictionaries that contain of approximately 35 000 simple and multi-word unit proper names (lemmas) that allow automatic generation of inflected forms. All produced word forms are supplied by detailed morpho-syntactic description. Besides that, all lemmas are equipped with various semantic codes appropriate to named entities. For full and precise recognition of named entities a large collection of finite state transducers is developed.

AlfaNum MD – AlfaNum Morphologic Dictionary of Serbian – was developed for the purpose of linguistic research and development of commercial applications of speech and other language technologies. It consists of 100,517 lexemes (3,888,407 inflected forms) with accentuation.

3.5.3. Speech databases

AlfaNum ASR - AlfaNum Speech Databases for ASR - consists of five speech databases: 1. W50S200 contains studio recordings of 200 speakers (53 words per speaker); 2. W150tf1000 contains recordings over the telephone channel (150 words, 600-1000 pronunciations per word); 3. AN_SPEAKER contains office recordings using computer microphone of 44 speakers (29 male, 15 female); 4. S70W100s120 contains recordings originally done in a studio using reel-to-reel tape recorder, later converted to digital format (120 speakers, 70 sentences + 100 isolated utterances per speaker); 5. AlfaNum SpeechDatII(E) contains recordings over the telephone channel (500 speakers, 50 utterances per speaker, according to the SpeechDatII standard).

AlfaNum TTS - AlfaNum Speech Databases for TTS - consists of five speech databases: 1. TTSlab2g2s contains corpus of diphones and dissylables isolated from logatomes and meaningful words recorded in a studio (961 diphone and 625 dissylables); 2. TTSlsMarina, 3. TTSlsSandra, and 4. TTSlsMarija represent corpora composed of a collection of meaningful texts and meaningless word sequences recorded in a studio (each of them comprising of approximately 2 hours of speech); 5. TTSlsSnezana represent corpora composed of a collection of meaningful texts recorded in a studio (Approximately 12 hours of speech).

3.5.4. Multimedia databases

DABI - Digital Archive of the Institute for Balkan Studies – was developed by the Institute for Balkan Studies from the Serbian Academy of Sciences and Arts. This database contains digitized audio (2000 hours of speech), video, photo and textual ethnographic material collected in Serbia in the scope of the field research related to Serbian and four minority languages.

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	annotationType
Corpus of Contemporary Serbian	SrpKor	korpus.matf.bg.ac.rs	corpus	113000000	word	PoS tagging
French-Serbian Aligned Corpus	SrpFranKor	korpus.matf.bg.ac.rs	corpus	1500000	word (in Serbian)	PoS tagging
English-Serbian Aligned Corpus	SrpEngKor	korpus.matf.bg.ac.rs	corpus	1000000	word (in Serbian)	PoS tagging; lemmatization (Serbian)
Serbian MSD Annotated Corpus	SrpLemKor	korpus.matf.bg.ac.rs	corpus	5000000	word	lemmatization; PoS; MSD
Multilingual Edition of Verne's Novel "Around the World in 80 Days"	Verne80days	korpus.matf.bg.ac.rs	corpus	71859	word (in French)	segmentation
Serbian Wordnet	SrpWN	korpus.matf.bg.ac.rs	lexicalConceptualResource	15200	synset (25579 words)	-
Serbian Morphological Dictionary	SrpRec	korpus.matf.bg.ac.rs	lexicalConceptualResource	96500	lemma	-
Serbian Named Entity Resources	SrpNER	korpus.matf.bg.ac.rs	lexicalConceptualResource	35000	lemma	-
AlfaNum Morphologic Dictionary of Serbian	AlfaNum MD	alfanum.ftn.uns.ac.rs	lexicalConceptualResource	100517	lemma	-
AlfaNum Text Corpus of Serbian	AlfaNumKor	alfanum.ftn.uns.ac.rs	corpus	200027	word	PoS, accentuation
AlfaNum Speech Databases for ASR	AlfaNum ASR	alfanum.ftn.uns.ac.rs	corpus	-	-	phoneme labels
AlfaNum Speech Databases for TTS	AlfaNum TTS	alfanum.ftn.uns.ac.rs	corpus	18	hour	phoneme labels
Digital Archive of the Institute for Balkan Studies	DABI	www.balkaninstitut.com	corpus	2000	hour	-
Corpus of Serbian Language	CSL	www.serbian-corpus.edu.rs/ns/eindex.htm	corpus	11000000	word	PoS; lemmatized
Corpus of Serbian Language	CSL	www.serbian-corpus.edu.rs/ns/eindex.htm	corpus	11000000	word	PoS; lemmatized

Table 3.5 Summary of Serbian language resources

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 still in development or are constantly updated, can be classified in the following categories:

3.6.1. Text corpora

Monolingual (Slovak) corpora

SNK – Slovak National Corpus a representative corpus of contemporary Slovak language written texts. SNK currently contains about 770 million words from broad variety of texts published since 1955 (1953 being the time of most recent substantial Slovak language orthography reform). The corpus is automatically lemmatised and MSD tagged. The documents are annotated with their genre, style and other bibliographic information. There are specialised subcorpora containing fiction, informational texts, professional texts, original Slovak fiction, texts written from 1955 to 1989, and a balanced subcorpus.

SK-WEB – Slovak Web Corpus contains texts downloaded from the .sk domain. The texts are automaticaly lemmatized and morphologically tagged.

Legal – Slovak Legal Texts Corpus contains the current (2011) body of Slovak Republic laws. The corpus has been prepared in collaboration with the Ministry of Justice of the Slovak Republic.

Bilingual and Multilingual (with Slovak as one language)

SK-EN – The Slovak-English Parallel Corpus original English fiction texts and their Slovak translations, with automatically aligned sentences.

SK-CS – The Slovak-Czech Parallel Corpus mostly fiction translated between Slovak and Czech (in both direction), with small amount of non-fiction texts and some translations from third language into both Czech and Slovak. The texts are automatically sentence-aligned, with some amount of texts aligned manually.

SK-RU - Slovak–Russian Parallel Corpus original Russian fiction texts and their Slovak translations, with automatically aligned sentences.

SK-FR - Slovak–French Parallel Corpus original French fiction texts and their Slovak translations, with automatically aligned sentences.

3.6.2. Lexical Conceptual Resources

MA - Slovak Morphological Lexicon full paradigms of 77000 lemmas, together with MSD tags, as used in the Slovak National Corpus. The lexicon serves as a basis for automatic morphological analysis and disambiguation.

WN - Slovak WordNet a network of lexical-semantic relations, an electronic thesaurus with a structure modelled on that of the Princeton WordNet. The WordNet describes the meaning of a lexical unit of one or more words by placing this unit in a network of links which represent such relations as synonymy, hypernyny, meronymy etc. The Slovak WordNet has been built semi-automatically, using information from bilingual Slovak-English

dictionary, and the synsets were then manually proofread. The Slovak synsets are mapped to equivalent English Princeton WordNet semantic equivalents, and contain translation into German, Polish and Lithuanian.

Slovak Treebank - Slovak language treebank consists of 50000 manually syntactically annotated sentences, using the Prague Dependency Treebank methodology (analytical level). Most of the sentences has been annotated by two independent annotators.

3.6.3. Speech databases

Hovor - Corpus of Spoken Slovak a corpus of sound recordings of different types of speech, with the emphasis on spontaneous speech. The recordings are transcribed orthographically and phonemically.

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	annotationType
Slovak National Corpus	prim	LSIL	corpus	770000000	token	PosTagging; MSD
Corpus of Spoken Slovak	hovor	LSIL				
Slovak Morphological Lexicon	ma	LSIL	lexicalConceptualResource	77000	lemma	MSD
Slovak Treebank	Slovak Treebank	LSIL	corpus; lexicalConceptualResource	50000	sentence	PosTagging; MSD; dependency analysis
Slovak WordNet	wn	LSIL	lexicalConceptualResource	12500	synset	
Slovak-English Parallel Corpus	sk-en	LSIL	corpus	1500000	sentence	PosTagging; MSD; alignment
Slovak-Czech Parallel Corpus	sk-cs	LSIL	corpus	700000	sentence	PosTagging; MSD; alignment
Slovak Web Corpus	Sk-web	LSIL	corpus	900000000	token	PosTagging; MSD
Slovak Legal Texts Corpus	legal	LSIL	corpus	146000000	token	PosTagging; MSD
Slovak-Russian Parallel Corpus	sk-ru	LSIL	corpus	100000	sentence	PosTagging; MSD; alignment
Slovak-French Parallel Corpus	sk-fr	LSIL	corpus	21000	sentence	PosTagging; MSD; alignment

Table 3.6 Summary of Slovak language resources

4. Appendices - detailed description of resources actually or potentially available

4.1. Appendix 1 - Bulgarian

resourceTitle	**Bulgarian National Corpus**
resourceName	BulNC
urlDownload	http://search.dcl.bas.bg
dateCreation	2008- ongoing project
projectPartner	IBL
IPRholder.organizationShortName	IBL
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
availability	available (pseudocorpus)
license	PUB / CC BY
resourceLocation	http://search.dcl.bas.bg
distributionMedium	internetBrowsing; web service
restrictionsOfUse	no
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	Construction of the Dictionary of Bulgarian Language
actualUse.actualUse	-
actualUse.useNLPspecific	-
description	The Bulgarian National Corpus is a large-scale, representative, publicly available corpus of Bulgarian. The BulNC can also be defined as a monolingual general corpus, fully morpho-syntactically (and partially semantically) annotated. Presently, the Bulgarian National corpus consists of about 450 000 000 tokens and includes more than 10 000 samples.
relevantPublications	Koeva Sv., D. Blagoeva and S. Kolkovska. Bulgarian National Corpus Project. – In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, D. Tapias (eds.), Valletta, European Language Resources Association (ELRA), 2010, pp. 3678-3684. ISBN 2-9517408-6-7
urlDocumentation	http://ibl.bas.bg/en/BGNC_search_en.htm
resourceType	corpus
resourceSubtype	reference corpus
mediaType	text
noLanguages	one

multilingualityType	-
languageId	BG
size	450 000 000
sizeUnit	token
annotationType	lemmatization; stemming; PosTagging

resourceTitle	**Bulgarian annotated corpora - Bulgarian PoS Annotated Corpus**
resourceName	BulPoSCor
urlDownload	http://search.dcl.bas.bg
dateCreation	2005 – 2010
projectPartner	IBL
IPRholder.organizationShortName	IBL
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
availability	available (pseudocorpus)
license	PUB / META-SHARE ACA NC
resourceLocation	http://search.dcl.bas.bg
distributionMedium	internetBrowsing; download; CD-ROM; DVD-R; bluRay; hardDisk; paperCopy; other
restrictionsOfUse	no
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	statistical tagger training
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	-
description	The POS Tagged Bulgarian Corpus is the result of the manual POS disambiguation of each wordform by language experts. The POS Tagged Bulgarian Corpus comprises more than 100 thousand words. It was designed through the extraction of samples of 300 words minimum (expanded to sentence boundary) from the Structured Brown Corpus of Bulgarian. Parts of the POS Tagged Corpus were used as training and test corpora in the creation of sevral Bulgarian taggers. The POS Tagged Bulgarian Corpus enables efficient search of language patterns and forms in the text.
relevantPublications	Koeva, Sv., Sv. Leseva, I. Stoyanova, E. Tarpomanova, M. Todorova, Bulgarian Tagged Corpora. In: Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages, 18-20 October 2006, Sofia, Bulgaria, pp. 78-86, 2006.
urlDocumentation	http://dcl.bas.bg/en/corpora_en.html
resourceType	corpus
resourceSubtype	annotated corpus

mediaType	text
noLanguages	one
multilingualityType	-
languageId	BG
size	150 000
sizeUnit	word
annotationType	lemmatization; stemming; PosTagging

resourceTitle	**Bulgarian annotated corpora - Bulgarian Sense Annotated Corpus**
resourceName	BulSemCor
urlDownload	http://dcl.bas.bg/semcor/bg/
dateCreation	2005 - 2010
projectPartner	IBL
IPRholder.organizationShortName	IBL
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
availability	available (pseudocorpus)
license	PUB / META-SHARE ACA NC
resourceLocation	http://search.dcl.bas.bg
distributionMedium	internetBrowsing; download; CD-ROM; DVD-R; bluRay; hardDisk; paperCopy; other
restrictionsOfUse	no
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	word sense disambiguation of Bulgarian
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	-

description	The Sense Tagged Bulgarian corpus (BulSemCor) contains sense-disambiguated lexical items defined in the context of occurrence (http://dcl.bas.bg/en/corpora_en.html). It consists of excerpts from the samples from the ""Brown"" Corpus of Bulgarian and preserves its overall structure: excerpts of minimum 100 words were clipped according to a methodology accounting for the concentration of the most frequent words from different classes. As a result the input corpus for semantic annotation amounts to app. 100 thousand words since the text excerpts were expanded to the left and right to sentence boundaries. Each lexical item (simple or compound word) in the Semantic Corpus is assigned manually the unique semantic or grammatical meaning from the Bulgarian wordnet which occurs in the particular context.The SemCor is used as a training and test set in the elaboration of a probability formalism for automatic word-sense disambiguation oriented towards machine translation.
relevantPublications	Koeva, Sv., Sv. Leseva, M. Todorova, Bulgarian Sense Tagged Corpus. In: Proceedings of the 5th SALTMIL Confernece on Minority Languages, Genoa, pp.79-87, 2006; Koeva, Sv., S. Lesseva, E. Tarpomanova, B. Rizov, Ts. Dimitrova and H. Kukova. Bulgarian Sense Annotated Corpus - Results and Achievements. In. Proceedings from the seventh international conference FASSBL, Dubtovnik, 41-49, 2010. ISSN 978-953-55375-2-6
urlDocumentation	http://dcl.bas.bg/en/corpora_en.html
resourceType	corpus
resourceSubtype	annotated corpus
mediaType	text
noLanguages	one
multilingualityType	-
languageId	BG
size	105 000
sizeUnit	word
annotationType	lemmatization; stemming; PosTagging
resourceTitle	**Bulgarian-X language parallel corpora**
resourceName	Bul-XCor
urlDownload	http://search.dcl.bas.bg
dateCreation	2010-ongoing project
projectPartner	IBL
IPRholder.organizationShortName	IBL
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
availability	available (pseudocorpus)
license	PUB / CC BY
resourceLocation	http://search.dcl.bas.bg
distributionMedium	internetBrowsing; web service
restrictionsOfUse	no

licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	machine translation
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	-
description	Most of the texts are Bulgarian translations of English originals but there are also translations from a third language into both English and Bulgarian. The most important condition for a text to be selected for the corpus was that it was either created after 1945 or its Bulgarian translation was made after that year. The fiction subcorpus consists of 340 samples texts per language – 34 553 474 words for the Bulgarian part and 39 590 472 words for the English part.
relevantPublications	-
urlDocumentation	-
resourceType	corpus
resourceSubtype	parallel corpus
mediaType	text
noLanguages	one
multilingualityType	parallel
languageId	BG
size	100 000 000
sizeUnit	token
annotationType	alignment; lemmatization; stemming; PosTagging

resourceTitle	**Bulgarian FrameNet**
resourceName	BulFrameNet
urlDownload	http://dcl.bas.bg/LexIt/
dateCreation	2001-ongoing project
projectPartner	IBL
IPRholder.organizationShortName	IBL
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg

availability	available-restricted use
license	ELDA / CC BY
resourceLocation	http://dcl.bas.bg/LexIt/
distributionMedium	internetBrowsing; ELDA
restrictionsOfUse	no
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	machine translation
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	-
description	The Bulgarian FrameNet (http://dcl.bas.bg/BulFrameNet.html) represents general semantic and language-specific lexical-semantic and syntactic combinatory properties of Bulgarian lexical units. The Bulgarian FrameNet (BulFrameNet) database so far contains unique descriptions of over 3,000 Bulgarian lexical units, approx. one tenth of them aligned with appropriate semantic frames, supports XML import and export and will be accessible, i.e., displayed and queried via the web. Each lexical entry consists of a lexical unit; a semantic frame from the English FrameNet, expressing abstract semantic structure; a grammatical class, defining the inflexional paradigm; a valency frame describing (some of) the syntactic and lexical-semantic combinatory properties (an optional component); and (semantically and syntactically) annotated examples.
relevantPublications	Koeva Sv. Lexicon and Grammar in Bulgarian FrameNet. – In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, D. Tapias (eds.), Valletta, European Language Resources Association (ELRA), 2010, pp. 3678-3684. ISBN 2-9517408-6-7; Koeva Sv. Integrating Semantic and Syntactic Annotation in Bulgarian FrameNet. In: Proceedings from the 29th International Conference on Lexis and Grammar, Belgrade, Serbia, September 15-18, 2010. ISBN 97886-758-080-5
urlDocumentation	-
resourceType	lexicalConceptualResource
resourceSubtype	framenet
mediaType	text
noLanguages	one
multilingualityType	-
languageId	BG
size	3 000
sizeUnit	lemma
annotationType	

resourceTitle	**Bulgarian Grammatical Dictionary**
resourceName	BulGram
urlDownload	http://dcl.bas.bg/est/dict.php
dateCreation	1999-2003
projectPartner	IBL
IPRholder.organizationShortName	IBL
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
availability	available-restricted use
license	PUB / ELDA
resourceLocation	http://dcl.bas.bg/est
distributionMedium	internetBrowsing; ELDA
restrictionsOfUse	no
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	various
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	-
description	The Grammar Dictionary of Bulgarian is an inflexion dictionary that consists of approximately 80 000 lemmas and allows automatic word form analysis and generation resulting in approximately 1 mln. 140 thousand word forms. The formal structure of the dictionary is based on finite state transducers (FSTs) that are widely applied in up-to- date electronic dictionaries. In its present state the Grammar Dictionary of Bulgarian represents in full the synthetic word formation patterns in Bulgarian and is therefore suitably transformed into spelling checking dictionary.
relevantPublications	Koeva, Sv. and M. Silberztein. Bulgarian and English Semantic Dictionaries for the Purposes of Information Retrieval, In: Computer Treatment of Slavic and East European Languages, ed. Radovan Garabik, Bratislava, Veda, pp. 193-203, 2005. ISBN 80-224-0895-6
urlDocumentation	-
resourceType	lexicalConceptualResource
resourceSubtype	lexicon
mediaType	text
noLanguages	one
multilingualityType	-
languageId	BG
size	85 000

sizeUnit	lemma
annotationType	

4.2. APPENDIX 2 – CROATIAN

resourceTitle	**Croatian National Corpus**
resourceName	HNK
urlDownload	http://hnk.ffzg.hr
dateCreation	ongoing work
projectPartner	FFZG
IPRholder.organizationShortName	FFZG
contact.Person.surname	Tadić
contact.Person.givenName	Marko
contact.Person.email	marko.tadic@ffzg.hr
availability	available-restricted use
license	PUB / CC BY-NC-SA
resourceLocation	FFZG
distributionMedium	internetBrowsing; webservice
restrictionsOfUse	attribution
licenseSignatory.Person.position	
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	balanced reference corpus of contemporary written Croatian standard language
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	linguistic evidence; frequency lists; n-grams production; language modelling; morphological processing; information extraction; text-mining; NERC, tree-banking; chunking; dependency parsing; document classification; machine translation
description	The Croatian National Corpus (HNK) is the largest balanced representative monolingual corpus for Croatian written standard language. It covers different domains, genres and topics starting with the topmost classification to prose fiction texts, faction texts and texts of mixed type. Its compilation starded in 1998, while the present version (v2.5) appeared in 2008. This version represents the 101.3 million tokens large, lemmatised and MSD tagged corpus of texts produced from 1990 up to present day. The HNK is encoded in XCES and tagging was performed by a hybrid tagger using the MULTEXT EAST compliant Croatian tagset. The HNK is stored on Manatee server and it is accessible freely for search using Bonito client. See http://hnk.ffzg.hr for further access details.
relevantPublications	Tadić, M. (2002) Building the Croatian National Corpus, LREC2002 Proceedings, Las Palmas, ELRA, Paris-Las Palmas, Vol. II, pp 441-446. Tadić, M. (2003) Jezične tehnologije i hrvatski jezik, Exlibris, Zagreb. Tadić, M. (2009) New version of the Croatian National Corpus. In: Hlaváčková, D.; Horák, A.; Osolsobě, K.; Rychlý, P. After Half a Century of Slavonic Natural Language Processing, Masaryk University, Brno, pp 199-205.
urlDocumentation	http://hnk.ffzg.hr
resourceType	corpus
resourceSubtype	reference corpus
mediaType	text
noLanguages	one
multilingualityType	
languageId	hrv
size	101,300,000
sizeUnit	token
annotationType	structural annotation; lemmatization; PoS Tagging

resourceTitle	**Croatian-English Parallel Corpus**
resourceName	HR-EN p-corp
urlDownload	
dateCreation	ongoing work
projectPartner	FFZG
IPRholder.organizationShortName	FFZG
contact.Person.surname	Tadić
contact.Person.givenName	Marko
contact.Person.email	marko.tadic@ffzg.hr
availability	available-restricted use
license	PUB / CC BY-NC-SA
resourceLocation	FFZG
distributionMedium	not yet available for internet access
restrictionsOfUse	academic-nonCommercialUse; commercialUse for a fee
licenseSignatory.Person.position	
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	bilingual lexicography; contrastive linguistic research; translation studies; machine translation
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	contrastive linguistic research; machine translation
description	The Croatian-English parallel corpus, which has been collected at the Institute of Linguistics at the Faculty of Humanities and Social Sciences, University of Zagreb is a bilingual parallel corpus based on the texts from the "Croatia Weekly" newspapers published in Zagreb from 1998 to 2000. This weekly newspaper digested texts on all important events and covered different domains such as politics, economy and finances, culture, tourism, sports, ecology etc. The corpus is 3.5 million tokens (1.9 en, 1.6 hr), sentence aligned and manually checked.
relevantPublications	Tadić, M. (2000) Building the Croatian-English Parallel Corpus, LREC2000 Proceedings, Athens, ELRA, Paris-Athens, Vol. I, pp 523-530. Tadić, M. (2003) Jezične tehnologije i hrvatski jezik, Exlibris, Zagreb
urlDocumentation	
resourceType	corpus
resourceSubtype	parallel corpus
mediaType	text
noLanguages	two
multilingualityType	parallel
languageId	hrv, eng
size	3,500,000
sizeUnit	token
annotationType	segmentation; alignment; structural annotation

resourceTitle	**Croatian Morphological Lexicon**
resourceName	HML
urlDownload	http://hml.ffzg.hr
dateCreation	ongoing work
projectPartner	FFZG
IPRholder.organizationShortName	FFZG
contact.Person.surname	Tadić
contact.Person.givenName	Marko
contact.Person.email	marko.tadic@ffzg.hr
availability	available-restricted use
license	PUB / CC BY-NC-SA
resourceLocation	FFZG
distributionMedium	internetBrowsing; webservice
restrictionsOfUse	academic-nonCommercialUse; commercialUse for a fee
licenseSignatory.Person.position	
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	lemmatization; PoS tagging; MSD tagging
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	lemmatization; PoS tagging; MSD tagging
description	The Croatian Morphological Lexicon is an inflectional lexicon which is compliant with MULTEXT EAST lexica format. It covers more than 100,000 lemmas: 45,000+ lemmas of general language, 15,000+ personal names and 50,000+ surnames registered in the Republic of Croatia. Number of generated word-forms is more than 4 million. This lexicon represents the resource background for Croatian Lemmatization Server (http://hml.ffzg.hr)
relevantPublications	Tadić, M.; Fulgosi, S. (2003) Building the Croatian Morphological Lexicon. In: Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages, Budapest, ACL, pp. 41-46. Tadić, M. (2005) The Croatian Lemmatization Server. Southern Journal of Linguistics, Vol. 29 (2005), 1-2, pp. 206-217 Bekavac, B.; Tadić, M. (2006) Inflectionally Sensitive Web Search in Croatian using Croatian Lemmatization Server. Proceedings of ITI2006 Conference, SRCE, Zagreb 2006, pp. 481-486.
urlDocumentation	http://hml.ffzg.hr
resourceType	lexicalResource
resourceSubtype	lexicon
mediaType	text
noLanguages	one
multilingualityType	
languageId	hrv
size	4,000,000
sizeUnit	lexical entry
annotationType	

resourceTitle	**Croatian Dependency Treebank**
resourceName	HOBS
urlDownload	http://hobs.ffzg.hr
dateCreation	ongoing work
projectPartner	FFZG
IPRholder.organizationShortName	FFZG
contact.Person.surname	Tadić
contact.Person.givenName	Marko
contact.Person.email	marko.tadic@ffzg.hr
availability	available-restricted use
license	CC BY-NC-SA
resourceLocation	FFZG
distributionMedium	not yet available for internet access
restrictionsOfUse	academic-nonCommercialUse; commercialUse for a fee
licenseSignatory.Person.position	
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	chunking; parsing; syntactic analysis
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	chunking; parsing; syntactic analysis
description	The Croatian Dependency Treebank is part of the Croatian National Corpus (i.e. Croatian part of the Croatian-English Parallel Corpus, CW2000) where ca 5000 sentences (ca 100,000 tokens) are manually annotated at the analytical layer following the Prague Dependency Treebank formalism adapted to Croatian. The treebank is being collected since 2007 and currently it's size is ca 4,700 sentences.
relevantPublications	Tadić, M. (2006) Croatian Dependency Treebank in Multilingual Context. Readings in Multilinguality: Selected papers for young researchers, Bulgarian Academy of Sciences, Sofia, pp. 125-128. Tadić, M. (2007) Building the Croatian Dependency Treebank: the initial stages. Suvremena lingvistika 63, (2007) pp. 85-92 Vučković, K.; Tadić, M.; Dovedan, Z. (2008) Rule Based Chunker for Croatian. LREC2008 Proceedings, Marrakesh, ELRA, Paris-Marrakesh.
urlDocumentation	http://hobs.ffzg.hr
resourceType	corpus
resourceSubtype	reference corpus
mediaType	text
noLanguages	one
multilingualityType	
languageId	hrv
size	4,700
sizeUnit	sentence
annotationType	segmentation

resourceTitle	**Croatian Vallency Dictionary**
resourceName	CROVALLEX
urlDownload	http://cal.ffzg.hr/crovallex
dateCreation	2008
projectPartner	FFZG
IPRholder.organizationShortName	FFZG
contact.Person.surname	Mikelić-Preradović
contact.Person.givenName	Nives
contact.Person.email	nmikelic@ffzg.hr
availability	available-restricted use
license	PUB / CC BY-NC-SA
resourceLocation	FFZG
distributionMedium	download
restrictionsOfUse	academic-nonCommercialUse; commercialUse for a fee
licenseSignatory.Person.position	
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	chunking; parsing; semantic role detection; NLP understanding
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	chunking; parsing; language generation
description	The Croatian Valency Lexicon (CROVALLEX) provides a formal description of valency frames of Croatian verbs. CROVALLEX 2.0008 was developed by Nives Mikelic Preradovic. The Functional Generative Description (FGD), being developed by Sgall and his collaborators since the 1960s, is used as the background theory in CROVALLEX 2.0008. for the description of valency frames of selected verbs. CROVALLEX 2.0008 contains 1740 verbs that were selected as the most frequent verbs from the Croatian frequency dictionary.
relevantPublications	Mikelić-Preradović, N. (2008) Approaches to the Development of the Machine Lexicon for Croatian Language, PhD thesis, University of Zagreb, Faculty of Humanities and Social Sciences Mikelić Preradović, N. (2010) CROVALLEX lexicon improvements: Subcategorization and semantic constraints. WSEAS transactions on computers, 9.
urlDocumentation	http://cal.ffzg.hr/crovallex
resourceType	lexicalResource
resourceSubtype	lexicon
mediaType	text
noLanguages	one
multilingualityType	
languageId	hrv
size	1740
sizeUnit	lexical entry
annotationType	

4.3. Appendix 3 – Hungarian

resourceTitle	**Hunglish parallel corpus**
resourceName	Hunglish
urlDownload	http://mokk.bme.hu/resources/hunglishc
dateCreation	2006
projectPartner	MOKK
IPRholder.organizationShortName	Center for Media Research and Education
contact.Person.surname	Dániel
contact.Person.givenName	Varga
contact.Person.email	daniel@mokk.bme.hu
availability	available-unrestricted use
license	CC BY
resourceLocation	hosted by owner
distributionMedium	internetBrowsing
restrictionsOfUse	attribution
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	machine translation; multilingual information retrieval
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	machine translation
description	The Hunglish Corpus is a free sentence-aligned Hungarian-English parallel corpus of about 2 million sentences. The corpus may be searched through a web-based sentence search service. This service has more than 200,000 visits per month.
relevantPublications	Varga Dániel, Németh László, Halácsy Péter, Kornai András, Trón Viktor. Parallel corpora for medium density languages. In the proceedings of the RANLP 2005. (A version of this paper appeared in: Nicolov et al eds: Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005. John Benjamins 2007.)
urlDocumentation	http://mokk.bme.hu/resources/hunglishcorpus http://hunglish.hu
resourceType	corpus
resourceSubtype	parallel corpus
mediaType	text
noLanguages	2
multilingualityType	parallel

languageId	hu; en
size	5 420 000
sizeUnit	token
annotationType	segmentation; alignment

resourceTitle	**Hungarian Wordnet**
resourceName	-
urlDownload	http://www.inf.u-szeged.hu/rgai/nlp?lang=en&page=nlpproj_hunont
dateCreation	2007
projectPartner	RILHAS
IPRholder.organizationShortName	Consortium of RILHAS, Szeged University and Morphologic Ltd.
contact.Person.surname	Váradi
contact.Person.givenName	Tamás
contact.Person.email	varadi@nytud.hu
availability	available soon (License construction is under discussion)
license	-
resourceLocation	hosted by the consortium
distributionMedium	CD-ROM
restrictionsOfUse	-
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP use
foreseenUse.useNLPspecific	information retrieval; word sense disambiguation; machine translation; coreference resolution; emotion detection; sentiment analysis
actualUse.actualUse	NLP applications
actualUse.useNLPspecific	coreference resolution; content analysis
description	The Hungarian WordNet is a multilingual ontology, meaning that most of its synsets were mapped to equivalent concepts in English (Princeton) WordNet v. 2.0. The ontology is also linked to entries of a Hungarian Monolingual explanatory dictionary and to the entries of the Hungarian verb valency frame lexicon.
relevantPublications	Miháltz, Márton, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, Tamás Váradi: Methods and Results of the Hungarian WordNet Project. In: Proceedings of The Fourth Global WordNet Conference, Szeged, Hungary (2008), pp. 311–321. Alexin, Zoltán, János Csirik, György Szarvas, András Kocsor, Márton Miháltz: Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction. In Proceedings of the Third International WordNet Conference (GWC 2006), Seogwipo, Jeju Island, Korea, January 22-26, 2006, pp. 291-292. Prószéky, Gábor, Miháltz Márton: Magyar WordNet: az első magyar lexikális szemantikai adatbázis. In: Magyar Terminológia 1 (2008) 1, pp. 43-57.
urlDocumentation	http://www.inf.u-szeged.hu/projectdirs/hlt/index_en.html
resourceType	lexicalConceptualResource
resourceSubtype	thesaurus, WordNet
mediaType	text
noLanguages	1
multilingualityType	-
languageId	hu

size	42 000
sizeUnit	synset
annotationType	WordNet relations, link to PWN synsets

resourceTitle	**Hungarian National Corpus**
resourceName	HNC
urlDownload	http://corpus.nytud.hu/mnsz/
dateCreation	2005
projectPartner	RILHAS
IPRholder.organizationShortName	RILHAS
contact.Person.surname	Váradi
contact.Person.givenName	Tamás
contact.Person.email	varadi@nytud.hu
availability	available-unrestricted use
license	CLARIN ACA+NC+Inf
resourceLocation	hosted by owner
distributionMedium	internetBrowsing
restrictionsOfUse	noDerivatives
licenseSignatory.Person.position	deputy-director
foreseenUse.foreseenUse	human use; NLP use
foreseenUse.useNLPspecific	word sense disambiguation; machine translation; named entity recognition;
actualUse.actualUse	human use, NLP use
actualUse.useNLPspecific	word sense disambiguation; machine translation; named entity recognition;
description	The Hungarian National Corpus is a the general purpose, representative corpus of today's written Hungarian language. It gives an exact, quantifiable picure of Hungarian language use. It includes bibliographical metadata, and encodes the boundaries of structural units (paragraphs, sentences). The corpus is automatically POS-tagged: we have lemma, part of speech, and morphological analysis for each word. The query interface is freely available to anyone. It incorporates Hungarian texts from Hungary, Slovakia, Transcarpathia, Transylvania and Vojvodina in five different genres: press, literature (from Digital Literature Academy), scientific, official and personal communication. The size of the corpus is currently 187 million words. The partially syncactically parsed form of the whole corpus can be queried by the "Verb Argument Browser" tool: http://corpus.nytud.hu/vab
relevantPublications	Váradi Tamás: The Hungarian National Corpus. In: Proceedings of the 3rd LREC Conference, Las Palmas, Spanyolország, 2002, 385-389 Sass Bálint: "Mazsola" - eszköz a magyar igék bővítményszerkezetének vizsgálatára. In: Váradi Tamás (szerk.): Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból, MTA Nyelvtudományi Intézet, Budapest, 2009
urlDocumentation	http://corpus.nytud.hu/mnsz/index_eng.html
resourceType	corpus
resourceSubtype	reference corpus
mediaType	text
noLanguages	1

multilingualityType	-
languageId	hu
size	187600000
sizeUnit	token
annotationType	MSD

resourceTitle	**Szeged NER corpus**
resourceName	-
urlDownload	http://www.inf.u-szeged.hu/rgai/nlp?lang=en&page=corpus_ne
dateCreation	2005
projectPartner	RILHAS
IPRholder.organizationShortName	University of Szeged
contact.Person.surname	Farkas
contact.Person.givenName	Richárd
contact.Person.email	rfarkas@inf.u-szeged.hu
availability	available - non-commercial use
license	CLARIN ACA NC
resourceLocation	hosted by owner
distributionMedium	download
restrictionsOfUse	non-commercial use
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	automated named entity recognition;
actualUse.actualUse	human use, NLP use
actualUse.useNLPspecific	information extraction, automated named entity recognition
description	The Szeged NER corpus is is a manually annotated part of the Szeged treebank, consisting of short business news. The used NER categories are (based on the CoNLL system (http://www.cnts.ua.ac.be/conll2003/ner/)) the following: PERSON, ORGANISATION, LOCATION and OTHER.
relevantPublications	Szarvas, György; Farkas, Richárd; Felföldi, László; Kocsor, András; Csirik, János 2006: A highly accurate Named Entity corpus for Hungarian. In: Electronic Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genova, Italy, 24-26 May. Farkas, Richárd; Szarvas, György; Kocsor, András 2006: Named Entity Recognition for Hungarian Using Various Machine Learning Algorithms. Acta Cybernetica 17(3): 633-646. Szarvas, György; Farkas, Richárd; Kocsor, András 2006: A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In: Discovery Science 2006, pp. 267-278.
urlDocumentation	http://www.inf.u-szeged.hu/rgai/nlp?lang=en&page=corpus_ne
resourceType	corpus
resourceSubtype	reference corpus

mediaType	text
noLanguages	1
multilingualityType	-
languageId	hu
size	220 000
sizeUnit	token
annotationType	POS tagging, NE tagging

resourceTitle	**Szeged corpus**
resourceName	-
urlDownload	http://www.inf.u-szeged.hu/projectdirs/hlt/en/Szeged%20Corpus%202.0_en.html
dateCreation	2005
projectPartner	RILHAS
IPRholder.organizationShortName	Consortium of University of Szeged, RILHAS, MorphoLogic Ltd.
contact.Person.surname	Vincze
contact.Person.givenName	Veronika
contact.Person.email	vinczev@inf.u-szeged.hu
availability	available-non-commercial use - usage is linked to a license agreement between the user and provider
license	CLARIN ACA NC
resourceLocation	hosted by owner
distributionMedium	CD-ROM; download
restrictionsOfUse	non-commercial use
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	morphosyntactic parsing, POS tagging
actualUse.actualUse	human use, NLP use
actualUse.useNLPspecific	information extraction, automated POS tagging
description	A morpho-syntactically annotated and manually disambiguated corpus of 1,2 million words.
relevantPublications	Alexin, Zoltán; Csirik, János; Gyimóthy, Tibor; Bibok, Károly; Hatvani, Csaba; Prószéky, Gábor; Tihanyi, László 2003: Manually Annotated Hungarian Corpus. In: Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics EACL'03, Budapest, Hungary, 15-17 April, pp. 53-56. Csendes, Dóra; Csirik, János; Gyimóthy, Tibor 2004: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In: Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora (LINC 2004) at The 20th International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland, 23-29 August, pp. 19-23.

urlDocumentation	http://www.inf.u-szeged.hu/projectdirs/hlt/en/Szeged%20Corpus%202.0_en.html
resourceType	corpus
resourceSubtype	reference corpus
mediaType	text
noLanguages	1
multilingualityType	-
languageId	hu
size	1200 000
sizeUnit	token
annotationType	POS tagging (MSD)

resourceTitle	**Szeged treebank**
resourceName	-
urlDownload	http://www.inf.u-szeged.hu/projectdirs/hlt/hu/szegedtreebank%202.0.html
dateCreation	2007
projectPartner	RILHAS
IPRholder.organizationShortName	Consortium of University of Szeged, RILHAS, MorphoLogic Ltd.
contact.Person.surname	Vincze
contact.Person.givenName	Veronika
contact.Person.email	vinczev@inf.u-szeged.hu
availability	available-non-commercial use - usage is linked to a license agreement between the user and provider
license	CLARIN ACA NC
resourceLocation	hosted by owner
distributionMedium	CD-ROM; download
restrictionsOfUse	non-commercial use
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	morphosyntactic parsing, POS tagging, syntactic parsing
actualUse.actualUse	human use, NLP use
actualUse.useNLPspecific	automated information extraction, semantic frame mapping, semantic role labeling, syntactic parsing
description	A manually checked treebank of 1,2 million words.
relevantPublications	Csendes, Dóra; Csirik, János; Gyimóthy, Tibor; Kocsor, András 2005: The Szeged Treebank. In: Matousek, Václav et al. (eds.): Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD 2005), Karlovy Vary, Czech Republic, September 12-16, 2005, Springer LNAI 3658, pp. 123-131.

urlDocumentation	http://www.inf.u-szeged.hu/projectdirs/hlt/en/Szeged%20Treebank%202.0_en.html
resourceType	corpus
resourceSubtype	syntax-corpus
mediaType	text
noLanguages	1
multilingualityType	-
languageId	hu
size	1200 000
sizeUnit	token
annotationType	POS tagging (MSD), treebank

resourceTitle	**Hungarian webcorpus**
resourceName	-
urlDownload	http://mokk.bme.hu/resources/ webcorpus
dateCreation	2003
projectPartner	MOKK
IPRholder.organizationShortName	Center for Media Research and Education
contact.Person.surname	Szakadát
contact.Person.givenName	István
contact.Person.email	i@syi.hu
availability	available-unrestricted use
license	CC-BY
resourceLocation	hosted by owner
distributionMedium	download
restrictionsOfUse	attribution
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	language modeling
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	language modeling
description	With over 1.48 billion words unfiltered (589m words fully filtered), this is by far the largest Hungarian language corpus, and it is available in its entirety under a permissive Open Content license. The Hungarian webcorpus was created as part of the WordSword project at the Media Research and Education Centre. The Webcorpus may be downloaded in two formats: as a frequency dictionary based on the texts and as the original texts.

relevantPublications	"Halácsy Péter, Kornai András, Németh László, Rung András, Szakadát István, Trón Viktor Creating open language resources for Hungarian In Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004), 2004. Kornai, A, Halácsy, P, Nagy, V, Oravecz, Cs, Trón, V, and Varga, D (2006). Web-based frequency dictionaries for medium density languages In: Proceedings of the 2nd International Workshop on Web as Corpus, edited by Adam Kilgarriff, Marco Baroni ACL-06, pages 1--9."
urlDocumentation	http://mokk.bme.hu/resources/webcorpus/
resourceType	corpus
resourceSubtype	reference corpus
mediaType	text
noLanguages	1
multilingualityType	-
languageId	hu
size	1 480 000 000
sizeUnit	token
annotationType	none

resourceTitle	**BABELHungarian Clear Speech Database**
resourceName	**BABEL**
urlDownload	http://alpha.tmit.bme.hu/speech/databases.php
dateCreation	2001
projectPartner	TMIT
IPRholder.organizationShortName	TMIT
contact.Person.surname	Vicsi
contact.Person.givenName	Klára
contact.Person.email	vicsi@tmit.bme.hu
availability	available-restricted use
license	ELRA
resourceLocation	hosted in international repository (ELRA)
distributionMedium	CD-ROM
restrictionsOfUse	-
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP use
foreseenUse.useNLPspecific	speech analysis; speech recognition; automatic speech recognition;
actualUse.actualUse	human use; NLP use
actualUse.useNLPspecific	speech analysis; speech recognition; automatic speech recognition;

description	BABEL database is an EUROM1 compatible database containing records from 60 speakers distributed in 3 speakers set (many, few, very few).
relevantPublications	-
urlDocumentation	http://alpha.tmit.bme.hu/speech/databases.php
resourceType	corpus
resourceSubtype	speech corpus
mediaType	audio
noLanguages	1
multilingualityType	-
languageId	hu
size	4
sizeUnit	hour
annotationType	annotation, segmentation

resourceTitle	**Hungarian Reference Speech Database**
resourceName	**MRBA**
urlDownload	http://alpha.tmit.bme.hu/speech/databases.php
dateCreation	2004
projectPartner	TMIT
IPRholder.organizationShortName	TMIT+ Szeged University
contact.Person.surname	Vicsi
contact.Person.givenName	Klára
contact.Person.email	vicsi@tmit.bme.hu
availability	available-restricted use
license	CLARIN RES
resourceLocation	hosted by the consortium
distributionMedium	DVD-R
restrictionsOfUse	-
licenseSignatory.Person.position	head of laboratory / dept
foreseenUse.foreseenUse	human use; NLP use
foreseenUse.useNLPspecific	speech analysis; speech recognition; automatic speech recognition;
actualUse.actualUse	human use; NLP use

actualUse.useNLPspecific	speech analysis; automatic speech recognition;
description	The database contains continuous read speech. During the planning of the corpus, we took into consideration the special characteristics of Hungarian language. Since the Hungarian is an agglutinative language, we needed to create a larger vocabulary in some categories, than it is mandatory. We tried to pay an extra attention to the topic 'phonetically rich sentences and words', to create a phonetically well balanced speech database for text independent speech recognizers. The database contains utterances read by 332 different speakers. The utterances were recorded in acoustically different locations.
relevantPublications	-
urlDocumentation	http://alpha.tmit.bme.hu/speech/databases.php
resourceType	corpus
resourceSubtype	speech corpus
mediaType	audio
noLanguages	1
multilingualityType	-
languageId	hu
size	6,5
sizeUnit	hour
annotationType	annotation, segmentation

resourceTitle	**Hungarian Telephone Speech Database**
resourceName	**MTBA**
urlDownload	http://alpha.tmit.bme.hu/speech/databases.php
dateCreation	2003
projectPartner	TMIT
IPRholder.organizationShortName	TMIT+ Szeged University
contact.Person.surname	Vicsi
contact.Person.givenName	Klára
contact.Person.email	vicsi@tmit.bme.hu
availability	available-restricted use
license	CLARIN RES
resourceLocation	hosted by the consortium
distributionMedium	DVD-R

restrictionsOfUse	-
licenseSignatory.Person.position	head of laboratory / dept
foreseenUse.foreseenUse	human use; NLP use
foreseenUse.useNLPspecific	speech analysis; speech recognition; automatic speech recognition;
actualUse.actualUse	human use; NLP use
actualUse.useNLPspecific	speech recognition; automatic speech recognition;
description	MTBA is a PSTN and mobil telephone voice Hungarian speech database. The database contains records based on the definition in SpeechDatE for the dialectical, age and sex balance and vocabulary. Important and different from the SpeechDatE database is, that the phonetically rich sentences and words have been segmented and labelled at phoneme level. The database has two parts. The first part (2 CDs) contains labeled application words, numbers, dates, spelling and names, the second part (1 CD) contains labeled and segmentated phonetically rich sentences and words.
relevantPublications	-
urlDocumentation	http://alpha.tmit.bme.hu/speech/databases.php
resourceType	corpus
resourceSubtype	speech corpus
mediaType	audio
noLanguages	1
multilingualityType	-
languageId	hu
size	5
sizeUnit	hour
annotationType	annotation, segmentation

resourceTitle	**Hungarian Telephone Client Speech**
resourceName	**MTÜBA**
urlDownload	-
dateCreation	2009
projectPartner	TMIT
IPRholder.organizationShortName	TMIT
contact.Person.surname	Vicsi
contact.Person.givenName	Klára
contact.Person.email	vicsi@tmit.bme.hu

availability	available-restricted use
license	CLARIN ACA+NC+ReD
resourceLocation	hosted by owner
distributionMedium	DVD-R
restrictionsOfUse	academic-nonCommercialUse; redistributionProhibited
licenseSignatory.Person.position	head of TMIT-LSA laboratory
foreseenUse.foreseenUse	human use; NLP use
foreseenUse.useNLPspecific	speech analysis; speech recognition; automatic speech recognition; spoken dialogue systems; emotion recognition; automatic person recognition; discourse analysis;
actualUse.actualUse	human use
actualUse.useNLPspecific	speech analysis; speech recognition; automatic speech recognition; emotion recognition; discourse analysis;
description	The MTÜBA database contains telephone calls recorded at the call centre of a service provider company. The database will be fully anonimized. The corpus consists of dialogues between the operator and the client. The ortographic transcription of the speech utterances is provided, clauses are segmentated (automatically followed by hand-correction). Parts of speech holding emotions acoustically audible are also labelled according to 4 basic emotions.
relevantPublications	-
urlDocumentation	-
resourceType	corpus
resourceSubtype	speech corpus
mediaType	audio
noLanguages	1
multilingualityType	-
languageId	hu
size	60
sizeUnit	hour
annotationType	annotation

resourceTitle	**Broadcast News Database**
resourceName	**BND**
urlDownload	-
dateCreation	2005
projectPartner	TMIT
IPRholder.organizationShortName	TMIT

contact.Person.surname	Vicsi
contact.Person.givenName	Klára
contact.Person.email	vicsi@tmit.bme.hu
availability	available-unrestricted use
license	CC BY SA NC
resourceLocation	hosted by owner
distributionMedium	download
restrictionsOfUse	academic-nonCommercialUse; shareAlike
licenseSignatory.Person.position	head of TMIT-LSA laboratory
foreseenUse.foreseenUse	human use; NLP use
foreseenUse.useNLPspecific	speech analysis; speech recognition; emotion recognition; automatic person recognition; discourse analysis;
actualUse.actualUse	human use
actualUse.useNLPspecific	speech analysis; speech recognition; emotion recognition; automatic person recognition; discourse analysis;
description	The Hungarian Broadcast News (HBN) database was collected as a member of the Broadcast News Interest Group of COST278, the COST action on Speech and Language Interaction in Telecommunications in cooperation of 10 different institutions throughout Europe. The Hungarian material consists of 3h and 30 minutes of recordings, transcribed and annotated, using the conventions of NIST (National Institute of Standards and Technology, USA).
relevantPublications	-
urlDocumentation	-
resourceType	corpus
resourceSubtype	multimedia and multimodal data
mediaType	audio; video
noLanguages	1
multilingualityType	-
languageId	hu
size	5
sizeUnit	hour
annotationType	annotation

resourceTitle	**Emotion Database**
resourceName	-
urlDownload	-

dateCreation	ongoing
projectPartner	TMIT
IPRholder.organizationShortName	TMIT
contact.Person.surname	Vicsi
contact.Person.givenName	Klára
contact.Person.email	vicsi@tmit.bme.hu
availability	available-restricted use
license	CLARIN ACA+ NC
resourceLocation	hosted by owner
distributionMedium	internet browsing
restrictionsOfUse	academic-nonCommercialUse;
licenseSignatory.Person.position	head of TMIT-LSA laboratory
foreseenUse.foreseenUse	human use
foreseenUse.useNLPspecific	emotion recognition
actualUse.actualUse	human use
actualUse.useNLPspecific	emotion recognition
description	Spoken databse holding emotionally reach utterances, labelling is done for emotions (8 basic emotions are labelled)
relevantPublications	-
urlDocumentation	-
resourceType	corpus
resourceSubtype	speech corpus
mediaType	audio
noLanguages	1
multilingualityType	-
languageId	hu
size	50
sizeUnit	hour
annotationType	annotation

resourceTitle	**Sound Gesture Database**
resourceName	-
urlDownload	-

dateCreation	ongoing
projectPartner	TMIT
IPRholder.organizationShortName	TMIT
contact.Person.surname	Vicsi
contact.Person.givenName	Klára
contact.Person.email	vicsi@tmit.bme.hu
availability	available-restricted use
license	CLARIN ACA+ NC
resourceLocation	hosted by owner
distributionMedium	internet browsing
restrictionsOfUse	academic-nonCommercialUse;
licenseSignatory.Person.position	head of TMIT-LSA laboratory
foreseenUse.foreseenUse	human use
foreseenUse.useNLPspecific	emotion recognition
actualUse.actualUse	human use
actualUse.useNLPspecific	emotion recognition
description	Database of sound gestures
relevantPublications	-
urlDocumentation	-
resourceType	lexicalConceptualResource
resourceSubtype	speech lexicon
mediaType	audio
noLanguages	1
multilingualityType	-
languageId	hu
size	770
sizeUnit	token
annotationType	annotation

resourceTitle	**Medical Database**
resourceName	-
urlDownload	-

dateCreation	ongoing
projectPartner	TMIT
IPRholder.organizationShortName	TMIT+ Semmelweis University
contact.Person.surname	Vicsi
contact.Person.givenName	Klára
contact.Person.email	vicsi@tmit.bme.hu
availability	available-restricted use
license	CLARIN RES
resourceLocation	hosted by the consortium
distributionMedium	DVD-R
restrictionsOfUse	-
licenseSignatory.Person.position	head of TMIT-LSA laboratory
foreseenUse.foreseenUse	human use
foreseenUse.useNLPspecific	speech recognition
actualUse.actualUse	human use
actualUse.useNLPspecific	-
description	The Medical database is a speech corpus holding utterances from persons suffering from different speech problems of organic origine.
relevantPublications	-
urlDocumentation	-
resourceType	corpus
resourceSubtype	speech corpus
mediaType	audio
noLanguages	1
multilingualityType	-
languageId	hu
size	1
sizeUnit	hour
annotationType	annotation

resourceTitle	**Broadcast Lectures Database**
resourceName	BLD or ME corpus (ME: Mindentudás Egyeteme)

urlDownload	http://mindentudas.hu/
dateCreation	ongoing
projectPartner	TMIT
IPRholder.organizationShortName	Hungarian Telecom, HAS
contact.Person.surname	Varga
contact.Person.givenName	Dániel
contact.Person.email	daniel@mokk.bme.hu
availability	availabe only for derivate works
license	to be clarified
resourceLocation	hosted by owner
distributionMedium	internet browsing
restrictionsOfUse	availabe only for derivate works
licenseSignatory.Person.position	Sallai László? (project managef of Mindentudás Egyeteme)
foreseenUse.foreseenUse	acoustic model training, speech recignition evaluation, etc.
foreseenUse.useNLPspecific	seech analysis;speaker identification; speaker verification; speech recognition; face verification; speech verification; face recognition; automatic speech recognition; automatic person recognition; multimedia development;l; information retrieval;
actualUse.actualUse	human use
actualUse.useNLPspecific	-
description	Recorded Broadcast Video Lectures from wide scientific topics for the public.
relevantPublications	-
urlDocumentation	-
resourceType	corpus
resourceSubtype	video corpus
mediaType	video
noLanguages	1
multilingualityType	-
languageId	hu
size	~150+
sizeUnit	hour
annotationType	annotation (?) transcription

resourceTitle	**Hungarian Speech Database of Holocaust Survovors' Testimonies**
resourceName	Hungarian MALACH (?)
urlDownload	-
dateCreation	2006
projectPartner	TMIT
IPRholder.organizationShortName	Spielberg's SOAH Foundation?, AITIA International
contact.Person.surname	Fegyó
contact.Person.givenName	Tibor
contact.Person.email	tfegyo@aitia.ai
availability	availabe only for derivate works
license	to be clarified
resourceLocation	AITIA International?
distributionMedium	DVD-R?
restrictionsOfUse	availabe only for derivate works
licenseSignatory.Person.position	Speech Research Director of Aitia International?
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	speech analysis;speaker identification; speaker verification; speech recognition; speech verification; automatic speech recognition; ;l; information retrieval;
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	-
description	Elderly people tell their history from the world war II, typically they are holocaust survivors. Originally videos were recorded with 2 microphones on two channel, but in Aitia International only 1 channel audio files are available with transcriptions.
relevantPublications	1 IEEE Transactions, 1 Springer LNCS, 1 Interspeech proc. Paper
urlDocumentation	-
resourceType	corpus
resourceSubtype	speech corpus
mediaType	audio
noLanguages	1
multilingualityType	-
languageId	hu
size	31
sizeUnit	hour

annotationType	annotation (?) transcription

resourceTitle	**Hungarian Parliamentary Speeches**
resourceName	
urlDownload	
dateCreation	ongoing
projectPartner	TMIT
IPRholder.organizationShortName	
contact.Person.surname	Fegyó
contact.Person.givenName	Tibor
contact.Person.email	tfegyo@aitia.ai
availability	availabe only for derivate works
license	to be clarified
resourceLocation	-
distributionMedium	-
restrictionsOfUse	availabe only for derivate works
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	speech analysis; speaker identification; speaker verification; speech recognition; speech verification; automatic speech recognition; ;l; information retrieval;
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	-
description	The Hungarian Parliamentary speeches are publicly available with approximate transcriptions. In this project, time alignment will be made and not alignable part will be marked.
relevantPublications	-
urlDocumentation	-
resourceType	corpus
resourceSubtype	speech corpus
mediaType	audio
noLanguages	1
multilingualityType	-
languageId	hu

size	1000+
sizeUnit	hour
annotationType	alignment, annotation

4.4. Appendix 4 - Polish

resourceTitle	**Polish Sejm Corpus**
resourceName	PSC
urlDownload	-
dateCreation	ongoing work
projectPartner	IPIPAN
IPRholder.organizationShortName	Polish Sejm
contact.Person.surname	Ogrodniczuk
contact.Person.givenName	Maciej
contact.Person.email	maciej.ogrodniczuk@ipipan.waw.pl
availability	available-unrestricted use
license	BSD 2-clause
resourceLocation	http:// clip.ipipan.waw.pl/PSC
distributionMedium	internetBrowsing/download
restrictionsOfUse	attribution
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	information retrieval
actualUse.actualUse	ongoing work
actualUse.useNLPspecific	ongoing work
description	The Polish Sejm Corpus will contain utterances of Polish Sejm members from cadencies 1-6. It will be morphologically annotated and made available for search with NKJP search tool – Poliqarp.
relevantPublications	-
urlDocumentation	-
resourceType	corpus
resourceSubtype	multimodal data
mediaType	text; audio; video
noLanguages	1
multilingualityType	-
languageId	pl
size	100 000 000

sizeUnit	word
annotationType	segmentation, lemmatization, POS, MSD, chunking, named entities

resourceTitle	**PoliMorf morphological dictionary**
resourceName	PoliMorf
urlDownload	-
dateCreation	ongoing work
projectPartner	IPIPAN
IPRholder.organizationShortName	IPIPAN
contact.Person.surname	Ogrodniczuk
contact.Person.givenName	Maciej
contact.Person.email	maciej.ogrodniczuk@ipipan.waw.pl
availability	available-unrestricted use
license	BSD 2-clause
resourceLocation	http:// clip.ipipan.waw.pl/PoliMorf
distributionMedium	internetBrowsing/download
restrictionsOfUse	attribution
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	morphosyntactic analysis
actualUse.actualUse	ongoing work
actualUse.useNLPspecific	ongoing work
description	PoliMorf combines two most important morphological dictionaries of Polish – Morfeusz SGJP and Morfologik. Morfeusz was developed by Zygmunt Saloni (author of linguistic data used in the analyser) and Marcin Woliński (programming part). The result morphological codes come from the IPI PAN Tagset (currently de facto standard for Polish) developed by Marcin Woliński and Adam Przepiórkowski for the annotation of the IPI PAN Corpus of Polish. Morfologik is based on current ispell dictionaries and Java libraries interfacing them. The result tags come from IPI Tagset.
relevantPublications	Marcin Woliński, Morfeusz — a Practical Tool for the Morphological Analysis of Polish, w: Mieczysław Kłopotek, Sławomir Wierzchoń, Krzysztof Trojanowski, red., Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings, s. 503–512, Springer, 2006.
urlDocumentation	http://sgjp.pl/morfeusz/index.html.en; http://morfologik.blogspot.com/
resourceType	lexicalConceptualResource
resourceSubtype	lexicon
mediaType	text

noLanguages	1
multilingualityType	-
languageId	pl
size	3000000
sizeUnit	word
annotationType	segmentation, lemmatization, POS, MSD

resourceTitle	**Polish Treebank**
resourceName	Treebank
urlDownload	-
dateCreation	ongoing work
projectPartner	IPIPAN
IPRholder.organizationShortName	IPIPAN
contact.Person.surname	Woliński
contact.Person.givenName	Marcin
contact.Person.email	wolinski@ipipan.waw.pl
availability	available-restricted use
license	GNU GPL
resourceLocation	http:// clip.ipipan.waw.pl/plTreebank
distributionMedium	internetBrowsing/download
restrictionsOfUse	shareAlike
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	parsing
actualUse.actualUse	ongoing work
actualUse.useNLPspecific	ongoing work
description	The treebank will contain trees created with Świgra – a deep parser of Polish implemented by Marcin Woliński on the basis of a metamorphosis grammar of Polish GFJP created by Świdziński.
relevantPublications	Marek Świdziński, Marcin Woliński, Towards a Bank of Constituent Parse Trees for Polish, w: Petr Sojka et al., red., Text, Speech and Dialogue, 13th International Conference, TSD 2010, Brno, September 2010, Proceedings, LNAI, tom 6231, s. 197–204, Springer, Heidelberg, 2010.
urlDocumentation	-
resourceType	corpus; lexicalConceptualResource
resourceSubtype	syntax corpus
mediaType	text

noLanguages	1
multilingualityType	-
languageId	pl
size	6000
sizeUnit	sentences
annotationType	segmentation, lemmatization, POS, MSD, chunking, deep parsing

resourceTitle	**1M National Corpus of Polish**
resourceName	1MNKJP
urlDownload	http://www.nkjp.pl/
dateCreation	June 2011
projectPartner	IPIPAN/ULodz
IPRholder.organizationShortName	IPIPAN for the corpus, various for the sources
contact.Person.surname	Przepiórkowski
contact.Person.givenName	Adam
contact.Person.email	adamp@ipipan.waw.pl
availability	available-restricted use
license	GNU GPL (1-million sample)
resourceLocation	http://www.nkjp.pl/
distributionMedium	internetBrowsing/download
restrictionsOfUse	restricted
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	various
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	various
description	The National Corpus of Polish is a shared initiative of four institutions: Institute of Computer Science at the Polish Academy of Sciences (coordinator), Institute of Polish Language at the Polish Academy of Sciences, Polish Scientific Publishers PWN, and the Department of Computational and Corpus Linguistics at the University of Łódź. It has been registered as a research-development project of the Ministry of Science and Higher Education. The list of sources for the corpus contains classic literature, daily newspapers, specialist periodicals and journals, transcripts of conversations, and a variety of short-lived and internet texts. The resources represent wide diversity with respect to the subject and genre. The spoken part covers both male and female speakers, in various age groups, coming from various regions in Poland.
relevantPublications	1. Katarzyna Głowińska, Adam Przepiórkowski. (2010). The Design of Syntactic Annotation Levels in the National Corpus of Polish. In: LREC 2010 proceedings. 2. Adam Przepiórkowski and Piotr Bański. (2009). XML Text Interchange Format in the National Corpus of Polish. To appear in the proceedings of PALC 2009. 3. Piotr Bański and Adam Przepiórkowski. (2009). Stand-off TEI Annotation: the Case of the National Corpus of Polish. In: the proceedings of the LAW-III workshop at ACL-IJCNLP 2009, Singapore, pp. 64-67.
urlDocumentation	http://www.nkjp.pl/

resourceType	corpus
resourceSubtype	reference corpus
mediaType	text
noLanguages	1
multilingualityType	-
languageId	pl
size	1M
sizeUnit	word
annotationType	segmentation, lemmatization, POS, MSD, chunking, word sense, named entities

resourceTitle	**Polish Parallel Corpora**
resourceName	PPC
urlDownload	-
dateCreation	ongoing work
projectPartner	ULodz /IPIPAN
IPRholder.organizationShortName	various
contact.Person.surname	Pęzik
contact.Person.givenName	Piotr
contact.Person.email	piotr.pezik@gmail.com
availability	available-unrestricted use
license	various (GNU GPL, CC BY-NC and public domain)
resourceLocation	http:// clip.ipipan.waw.pl/plPPC
distributionMedium	download
restrictionsOfUse	attribution
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	translation
actualUse.actualUse	ongoing work
actualUse.useNLPspecific	ongoing work
description	The corpus will include various parallel resources, including Acquis Communautaire, Europarl, OPUS (EMEA, KDE, Open Subtitles), Parasol, CORDIS and RAPID news etc.
relevantPublications	-
urlDocumentation	-
resourceType	corpus
resourceSubtype	parallel corpus
mediaType	text
noLanguages	5
multilingualityType	parallel
languageId	pl;en;de;es;fr
size	2 000 000
sizeUnit	word
annotationType	segmentation

resourceTitle	**Polish Spoken Multimedia Corpus**
resourceName	SMC
urlDownload	-
dateCreation	ongoing work
projectPartner	IPIPAN
IPRholder.organizationShortName	IPIPAN
contact.Person.surname	Marciniak
contact.Person.givenName	Małgorzata
contact.Person.email	mm@ipipan.waw.pl
availability	available-restricted use
license	CC BY-NC
resourceLocation	http:// clip.ipipan.waw.pl/plSMC
distributionMedium	download
restrictionsOfUse	attribution
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	speech analysis; discourse analysis; spoken dialogue systems
actualUse.actualUse	NLP applications
actualUse.useNLPspecific	speech analysis; spoken dialogue systems
description	The corpus contains transcripts of spontaneous informal conversations with the original recording files. It is available as TEI P5 encoded spoken Polish data.
relevantPublications	"1. Agnieszka Mykowiecka, Małgorzata Marciniak and Katarzyna Głowińska. ""Automatic Semantic Annotation of Polish Dialogue Corpus"". TSD 2008, September 2008. 2. Agnieszka Mykowiecka, Krzysztof Marasek, Małgorzata Marciniak, Joanna Rabiega-Wiśniewska, Ryszard Gubrynowicz. ""On Construction of Polish Spoken Dialogs Corpus, The 2nd Linguistic Annotation Workshop (The LAW II)"" held in conjunction with LREC 2008, Marrakech, Morocco, May 26-28, 2008."
urlDocumentation	-
resourceType	corpus
resourceSubtype	speech corpus
mediaType	text; audio
noLanguages	1
multilingualityType	-
languageId	pl
size	2 500 000
sizeUnit	words (approx. 1000 conversations)
annotationType	segmentation, POS

resourceTitle	**Polish Spoken Conversational Corpus**

resourceName	PSC C
urlDownload	http://pelcra.pl/?page_id=3
dateCreation	ongoing work
projectPartner	ULodz
IPRholder.organizationShortName	ULodz, NKJP
contact.Person.surname	Pęzik
contact.Person.givenName	Piotr
contact.Person.email	piotr.pezik@gmail.com
availability	available-restricted use
license	GNU GPL
resourceLocation	http://www.nkjp.uni.lodz.pl/spoken.jsp
distributionMedium	http://www.nkjp.uni.lodz.pl/spoken.jsp
restrictionsOfUse	shareAlike, attribution
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	representation of casual conversational Polish
actualUse.actualUse	speech research
actualUse.useNLPspecific	NKJP/Utrecht Institute of Linguistics parsing phonological structures
description	The Polish Spoken Conversational Corpus was originally a sub-corpus of the spoken part of the PELCRA Reference Corpus of Polish, which has been further expanded withing the National Corpus of Polish project. The corpus is composed of spontaneous natural conversations. After obtaining the participants' permission, the recordings were carefully transcribed and annotated with information about the respondents' sex, age, and educational level. All the original recordings have been digitized so that the conversation may not only be read, but also heard. At present the corpus contains some 1 800,000 words of transcribed speech accompanied by 75 hours of digitized recordings.
relevantPublications	Piotr Pęzik. Język mówiony w NKJP. Podręcznik użytkownika NKJP. PWN, 2011.
urlDocumentation	http://nkjp.uni.lodz.pl/spoken.jsp
resourceType	corpus
resourceSubtype	speech corpus
mediaType	text; audio
noLanguages	1
multilingualityType	-
languageId	pl
size	1 500 000
sizeUnit	word
annotationType	segmentation, discourse, sociolinguistic

resourceTitle	**The corpus of frequency dictionary of Polish language of the XX century sixties**
resourceName	WKSF
urlDownload	http://clip.ipipan.waw.pl/Polish%20language%20of%20the%20XX%20century%20sixties
dateCreation	ongoing work
projectPartner	IPIPAN
IPRholder.organizationShortName	Ida Kurcz, Andrzej Lewicki, Jadwiga Sambor, Krzysztof Szafran and Jerzy Woronczak
contact.Person.surname	Ogrodniczuk
contact.Person.givenName	Maciej
contact.Person.email	maciej.ogrodniczuk@ipipan.waw.pl
availability	available-restricted use
license	GNU GPL
resourceLocation	http://clip.ipipan.waw.pl/PL196x
distributionMedium	download
restrictionsOfUse	shareAlike
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	statistical tagger training
actualUse.actualUse	NLP applications
actualUse.useNLPspecific	statistical tagger training
description	The corpus was collected for the purpose of creating general frequency dictionary of contemporary Polish. The work started in 1967. Partial results were published between 1972 and 1977, the completed dictionary in 1990. The corpus was later augmented in various respects, both by manual editing and automated procedures. Corpus data contain 10,000 samples divided into 5 parts: essays, news, scientific texts, fiction and plays. Every sample is approximately 50 words long, they all come from texts published between 1963 and 1967 and contain bibliographic description of its source. Each word is tagged with its base form and some morphological properties. Sentence boundaries are also marked. Currently the corpus undergoes manual verification of the morphological descriptions.
relevantPublications	1. Kurcz, Ida; Lewicki, Andrzej; Sambor, Jadwiga; Szafran, Krzysztof; Woronczak, Jerzy. Słownik frekwencyjny polszczyzny współczesnej. (In Polish, EN: Frequency dictionary of contemporary Polish). Kraków, 1990. Institute of Polish Philology, Polish Academy of Sciences. 2. Bień, Janusz S.; Woliński, Marcin. Wzbogacony korpus Słownika frekwencyjnego polszczyzny współczesnej. (In Polish, EN: Enhanced corpus of the Frequency dictionary of contemporary Polish). [In:] Prace lingwistyczne dedykowane prof. Jadwidze Sambor. Jadwiga Linde-Usiekniewicz (ed.), pp. 6-10, Warszawa 2003, Faculty of Polish Philology, Warsaw University.
urlDocumentation	http://clip.ipipan.waw.pl/Polish%20language%20of%20the%20XX%20century%20sixties
resourceType	corpus
resourceSubtype	reference corpus
mediaType	text
noLanguages	1
multilingualityType	-
languageId	pl

size	500 000
sizeUnit	word
annotationType	segmentation, lemmatization, POS, MSD

resourceTitle	**NE resources with gazetteers**
resourceName	plNER
urlDownload	-
dateCreation	ongoing work
projectPartner	IPIPAN
IPRholder.organizationShortName	IPIPAN/University Tours
contact.Person.surname	Savary
contact.Person.givenName	Agata
contact.Person.email	agata.savary@univ-tours.fr
availability	available-restricted use
license	GNU GPL
resourceLocation	http://www.clip.waw.pl/plNER
distributionMedium	internetBrowsing/download
restrictionsOfUse	shareAlike
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	named entity recognition
actualUse.actualUse	NLP applications
actualUse.useNLPspecific	information extraction
description	The gazetters has been obtained from existing sources and supplemented with additional language-specific resources acquired from the Web. Whenever appropriate, inflected forms were generated using Morfeusz SGJP generator. Gazetteer data was used in the process of named entity annotation of NKJP.
relevantPublications	1. Małgorzata Marciniak, Joanna Rabiega-Wiśniewska, Agata Savary, Marcin Woliński, Celina Heliasz, Constructing an Electronic Dictionary of Polish Urban Proper Names, w: Mieczysław A. Kłopotek, Adam Przepiórkowski, Sławomir T. Wierzchoń, Krzysztof Trojanowski, red., Recent Advances in Intelligent Information Systems, Challenging Problems of Science, EXIT, Warsaw, 2009, ISBN 978-83-60434-59-8, s. 233–246. 2. Agata Savary, Joanna Rabiega-Wiśniewska, Marcin Woliński, Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex, w: Aspects of Natural Language Processing, LNCS 5070, s. 111–141, Springer, 2009.
urlDocumentation	-
resourceType	lexicalConceptualResource
resourceSubtype	lexicon
mediaType	text
noLanguages	1
multilingualityType	-
languageId	pl

size	229681
sizeUnit	word forms
annotationType	-

resourceTitle	**Polish WordNet (Słowosieć)**
resourceName	plWordNet
urlDownload	http://plwordnet.pwr.wroc.pl/wordnet/
dateCreation	ongoing work
projectPartner	IPIPAN
IPRholder.organizationShortName	Wrocław University of Technology
contact.Person.surname	Piasecki
contact.Person.givenName	Maciej
contact.Person.email	maciej.piasecki@pwr.wroc.pl
availability	available-restricted use
license	free for non-commercial use; license can be obtained on request
resourceLocation	http://plwordnet.pwr.wroc.pl/wordnet/
distributionMedium	internetBrowsing/download
restrictionsOfUse	academic-nonCommercialUse; commercial use possible
licenseSignatory.Person.position	Prof. Eugeniusz Rusiński, Wrocław University of Technology
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	various
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	various (e.g. coreference resolution)
description	Polish WordNet is a network of lexical-semantic relations, an electronic thesaurus with a structure modelled on that of the Princeton WordNet and those constructed in the EuroWordNet project. Polish WordNet describes the meaning of a lexical unit of one or more words by placing this unit in a network of links which represent such relations as synonymy, hypernyny, meronymy etc. To reduce the cost of the project, Polish WordNet has been built semi-automatically. Lexical relations were automatically recognized in large corpora of Polish and suggested to linguists/lexicographers via a graphical interface.
relevantPublications	Maciej Piasecki, Stanisław Szpakowicz, Bartosz Broda. A Wordnet from the Ground Up.
urlDocumentation	http://plwordnet.pwr.wroc.pl/wordnet/
resourceType	lexicalConceptualResource
resourceSubtype	wordnet
mediaType	text
noLanguages	1
multilingualityType	-
languageId	pl
size	20223

sizeUnit	lemmas (17695 synsets)
annotationType	-

resourceTitle	**Polish Valency Dictionary**
resourceName	PVD
urlDownload	http://clip.ipipan.waw.pl/PVD
dateCreation	ongoing work
projectPartner	IPIPAN
IPRholder.organizationShortName	IPIPAN, Marek Świdziński
contact.Person.surname	Ogrodniczuk
contact.Person.givenName	Maciej
contact.Person.email	maciej.ogrodniczuk@ipipan.waw.pl
availability	available-unrestricted use
license	FreeBSD or CC-BY-NC (expected)
resourceLocation	http://clip.ipipan.waw.pl/PVD
distributionMedium	download
restrictionsOfUse	attribution (expected)
licenseSignatory.Person.position	
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	deep parsing, corpus annotation
actualUse.actualUse	ongoing work
actualUse.useNLPspecific	ongoing work
description	The Polish Valency Dictionary is a merger of several valency dictionaries that have been made available in 2009 (a national Ministry of Science and Higher Education project "Automatic detection of semantic dependencies within verb argument structures in large treebanks") and in 2011 (a national Ministry of Science and Higher Education research grant "Construction of a treebank for Polish using automatic syntactic analysis") with an older, but very popular dictionary by prof. Świdziński.
relevantPublications	-
urlDocumentation	-
resourceType	lexicalConceptualResource
resourceSubtype	-
mediaType	text
noLanguages	1
multilingualityType	-
languageId	pl
size	1500
sizeUnit	entries
annotationType	valency annotation

resourceTitle	**Cross-lingual Repository of Named Entities**
resourceName	CRON
urlDownload	http://clip.ipipan.waw.pl/CRON
dateCreation	ongoing work
projectPartner	IPIPAN
IPRholder.organizationShortName	IPIPAN, University of Tours
contact.Person.surname	Ogrodniczuk
contact.Person.givenName	Maciej
contact.Person.email	maciej.ogrodniczuk@ipipan.waw.pl
availability	available-unrestricted use
license	FreeBSD, CC-BY-NC or GPL (expected)
resourceLocation	http://clip.ipipan.waw.pl/CRON
distributionMedium	download
restrictionsOfUse	attribution (expected)
licenseSignatory.Person.position	
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	corpus annotation, multilingual language processing
actualUse.actualUse	ongoing work
actualUse.useNLPspecific	ongoing work
description	A new multilingual dictionary of named entities prepared with co-operation of the University of Tours.
relevantPublications	-
urlDocumentation	-
resourceType	lexicalConceptualResource
resourceSubtype	-
mediaType	text
noLanguages	1
multilingualityType	-
languageId	pl
size	10000
sizeUnit	entries
annotationType	LMF

4.5. Appendix 5 - Serbian

resourceTitle	**Corpus of Contemporary Serbian**
resourceName	SrpKor
urlDownload	nlp.matf.bg.ac.rs
dateCreation	2002-
projectPartner	UBG
IPRholder.organizationShortName	UBG-MATF
contact.Person.surname	Vitas
contact.Person.givenName	Dusko
contact.Person.email	vitas@matf.bg.ac.rs
availability	available-restricted use
license	CC NC BY
resourceLocation	korpus.matf.bg.ac.rs
distributionMedium	internetBrowsing
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Dusko Vitas
foreseenUse.foreseenUse	human use
foreseenUse.useNLPspecific	various
actualUse.actualUse	human use
actualUse.useNLPspecific	various
description	The Corpus of Contemporary Serbian was initially constructed in cooperation of three institutions: Faculty of Mathematics (coordinator) and Faculty of Philology, University of Belgrade and Faculty of Philosophy, University of Novi Sad. It was partly funded by the Ministry of Education and Science. The list of sources for the corpus contains literature, monographs, daily newspapers, specialist periodicals and journals, manuals and textbooks. The resources represent wide diversity with respect to the subject and genre.
relevantPublications	1. Cvetana Krstev, Duško Vitas, “Corpus and Lexicon - Mutual Incompletness”, in Proceedings of the Corpus Linguistics Conference, 14-17 July 2005, Birmingham, eds. Pernilla Danielsson and Martijn Wagenmakers, ISSN 1747-9398
urlDocumentation	korpus.matf.bg.ac.rs/prezentacija
resourceType	corpus
resourceSubtype	reference corpus
mediaType	text
noLanguages	1
multilingualityType	-
languageId	SR
size	113000000
sizeUnit	word
annotationType	segmentation; PoS tagging

resourceTitle	**French-Serbian Aligned Corpus**
resourceName	SrpFranKor
urlDownload	nlp.matf.bg.ac.rs
dateCreation	1999-
projectPartner	UBG
IPRholder.organizationShortName	UBG-MATF
contact.Person.surname	Vitas
contact.Person.givenName	Dusko
contact.Person.email	vitas@matf.bg.ac.rs
availability	available-restricted use
license	CC NC BY
resourceLocation	korpus.matf.bg.ac.rs
distributionMedium	CD-ROM; internetBrowsing
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Dusko Vitas
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	various
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	various
description	The corpus includes French or Serbian source literary texts and their translations. Texts are segment aligned and manually checked to obtain one-to-one alignment.
relevantPublications	1. Duško Vitas, Cvetana Krstev, “Literature and Aligned Texts”, in Readings in Multilinguality, eds. Milena Slavcheva, Galia Angelova and Kiril Simov, pp. 148-155, Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria, 2006.; 2. Duško Vitas, Cvetana Krstev, Eric Laporte, “Preparation and exploitation of Bilingual Texts”, in Lux Coreana, No. 1, pp. 110-132, Han-Seine, 2006.; 3. Duško Vitas, Cvetana Krstev, “Structural derivation and meaning extraction: a comparative study on French-Serbo-Croatian parallel texts”, in Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora, eds. Geoff Barnbrook, Pernilla Danielsson, Michaela Mahlberg, pp. 166-178, The University of Birmingham Press, Birmingham, 2005.
urlDocumentation	korpus.matf.bg.ac.rs/prezentacija
resourceType	corpus
resourceSubtype	parallel corpus
mediaType	text
noLanguages	2
multilingualityType	parallel
languageId	FR; SR

size	1500 000
sizeUnit	word (in Serbian)
annotationType	segmentation; PoS tagging

resourceTitle	**English-Serbian Aligned Corpus**
resourceName	SrpEngKor
urlDownload	nlp.matf.bg.ac.rs
dateCreation	2004-
projectPartner	UBG
IPRholder.organizationShortName	UBG-MATF
contact.Person.surname	Krstev
contact.Person.givenName	Cvetana
contact.Person.email	cvetana@matf.bg.ac.rs
availability	available-restricted use
license	CC NC BY
resourceLocation	korpus.matf.bg.ac.rs
distributionMedium	internetBrowsing
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Dusko Vitas
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	various
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	various
description	The corpus includes English or Serbian source texts from the domains: education, health, legislation, jurisprudence. Texts are segment and lemmatized and PoS tagged, aligned and manually checked.
relevantPublications	1. Zoran Popović, Taggers applied on texts in Serbian, Infotheca, Vol. XI (2), Belgrade, 2010
urlDocumentation	korpus.matf.bg.ac.rs/prezentacija
resourceType	corpus
resourceSubtype	parallel corpus
mediaType	text
noLanguages	2
multilingualityType	parallel
languageId	EN; SR
size	1 000 000
sizeUnit	word (in Serbian)
annotationType	segmentation; PoS tagging; lemmatization (Serbian)

resourceTitle	**Serbian MSD Annotated Corpus**
resourceName	SrpLemKor
urlDownload	nlp.matf.bg.ac.rs
dateCreation	2011-
projectPartner	UBG
IPRholder.organizationShortName	UBG-MATF
contact.Person.surname	Utvić
contact.Person.givenName	Miloš
contact.Person.email	misko@matf.bg.ac.rs
availability	available-restricted use
license	CC NC BY
resourceLocation	korpus.matf.bg.ac.rs
distributionMedium	internetBrowsing
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Dusko Vitas
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	various
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	various
description	The Serbian MSD Annotated Corpus will consist of a sample of various texts from SrpKor. It will be lemmatized and MSD tagged.
relevantPublications	-
urlDocumentation	korpus.matf.bg.ac.rs/prezentacija
resourceType	corpus
resourceSubtype	corpus
mediaType	text
noLanguages	1
multilingualityType	-
languageId	SR
size	5 000 000
sizeUnit	word
annotationType	segmentation; lemmatization; PoS; MSD

resourceTitle	**Multilingual Edition of Verne's Novel "Around the World in 80 Days"**
resourceName	Verne80days
urlDownload	nlp.matf.bg.ac.rs
dateCreation	1998-
projectPartner	UBG
IPRholder.organizationShortName	UBG-MATF
contact.Person.surname	Vitas
contact.Person.givenName	Dusko
contact.Person.email	vitas@matf.bg.ac.rs
availability	available-restricted use
license	CC NC BY
resourceLocation	korpus.matf.bg.ac.rs
distributionMedium	DVD-R
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Dusko Vitas
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	various
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	various
description	This edition contains 25 translations of Jules Verne's novel "Around the World in 80 Days". Presently 15 of these translations are one-to-one segment aligned with the source text, and the alignment is manually checked. All other translations will also be aligned.
relevantPublications	1. Duško Vitas, Svetla Koeva, Cvetana Krstev, Ivan Obradović, “Tour du monde through the dictionaries”, Actes du 27eme Colloque International sur le Lexique et la Gammaire, L'Aquila, 10-13 septembre 2008, eds. M. Constant, T, Nakamura, M. De Gioia, S. Vecchiato, pp.249-256, Universite Paris-Est, Institut Gaspard-Monge, 2008.; 2. Emeline Lecuit, Denis Maurel, Duško Vitas, Cvetana Krstev, “Temporal Expressions: Comparisons in a Multilingual Corpus”, in Proceedings of 4th Language & Technology Conference, November 6-8, 2009, Poznań, Poland, ed. Zygmunt Vetulani, IMPRESJA Widawnictwa Elektroniczne S.A., Poznań, 2009.
urlDocumentation	korpus.matf.bg.ac.rs/prezentacija
resourceType	corpus
resourceSubtype	parallel corpus
mediaType	text
noLanguages	25
multilingualityType	parallel
languageId	FR; SR; EN; DE; PL; BG; HU; CR; SI; SK; RO; RU; GR; AL; PT; ES; IT; TR; CN; NL…
size	71859
sizeUnit	word (in French)

annotationType	segmentation

resourceTitle	**Serbian Wordnet**
resourceName	SrpWN
urlDownload	nlp.matf.bg.ac.rs
dateCreation	2002-
projectPartner	UBG
IPRholder.organizationShortName	UBG-MATF
contact.Person.surname	Pavlović-Lažetić
contact.Person.givenName	Gordana
contact.Person.email	gordana@matf.bg.ac.rs
availability	available-restricted use
license	CC NC BY
resourceLocation	korpus.matf.bg.ac.rs
distributionMedium	internetBrowsing; DVD-R
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Gordana Pavlović-Lažetić
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	various
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	various
description	Serbian WordNet (SrpWN) represents a lexical semantic network, containing synsets with glosses and various semantic relations, such as antonymy, meronymy, causation, category domain, etc. Currently the Serbian Wordnet contains near to 16,000 synsets. The initial version of the Serbian Wordnet was produced in the scope of the EU-funded Balkanet project. Through interlingual relations it is connected to English Wordnet,and wordnets of many other languages.
relevantPublications	1.Ivan Obradović, Cvetana Krstev, Gordana Pavlović-Lažetić, Duško Vitas, "Corpus Based Validation of WordNet Using Frequency Parameters", in Proceedings of the GWC : Second International WordNet Conference, Brno, Czech Republic, January 20-23, 2004, eds. P. Sojka, K. Pala, P. Smrž, Ch. Fellbaum, P. Vossen, ed. 1, pp. 181-186, Masaryk University, Brno, 2004.; 2.Svetla Koeva, Cvetana Krstev, Duško Vitas, "Morpho-semantic Relations in WordNet - a Case Study for two Slavic Languages", In the Proceedings of Global WordNet Conference 2008, eds. Attila Tanacs et al, University of Szeged, Department of Informatics, pp. 239-253, 2008.;3. Cvetana Krstev, Ivan Obradović, Duško Vitas, "An Approach to the Development of Language Specific Concepts in Wordnets", In Southern Journal of Linguistics, Special Theme: South Slavic and Balkan Languages, Mila Dimitrova-Vulchanova (ed.), Vo. 29, No. 1/2, pp. 106-118, Department of Modern Linguistics, University of Mississippi, 2008.
urlDocumentation	korpus.matf.bg.ac.rs/prezentacija
resourceType	lexicalConceptualResource
resourceSubtype	wordnet
mediaType	text

noLanguages	1
multilingualityType	-
languageId	SR
size	15 200
sizeUnit	synset (25579 words)
annotationType	-

resourceTitle	**Serbian Morphological Dictionary**
resourceName	SrpRec
urlDownload	nlp.matf.bg.ac.rs
dateCreation	2000-
projectPartner	UBG
IPRholder.organizationShortName	UBG-MATF
contact.Person.surname	Krstev
contact.Person.givenName	Cvetana
contact.Person.email	cvetana@matf.bg.ac.rs
availability	available-restricted use
license	CC NC BY
resourceLocation	korpus.matf.bg.ac.rs
distributionMedium	DVD-R
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Dusko Vitas
foreseenUse.foreseenUse	NLP applications
foreseenUse.useNLPspecific	various
actualUse.actualUse	NLP applications
actualUse.useNLPspecific	various
description	The Morphological Dictionary of Serbian is a dictionary that consists of approximately 90 000 simple lemmas and 6,500 multi-word lemmas that allows automatic generation of inflected forms using finite state transducers (approximately 4 million simple word forms). All produced word forms are supplied by detailed morpholo-syntactic description. Besides that, all lemmas are equipped with semantic, syntactic, domain and other codes.
relevantPublications	1. Cvetana Krstev, Duško Vitas, Gordana Pavlović-Lažetić, “Resources and Methods in the Morphosyntactic Processing of Serbo-Croatian”, In Formal Description of Slavic Languages: The Fifth Conference, Leipzig 2003, Zybatow, Gerhild et al. (eds.), Peter Lang: Frankfurt am Main, pp. 3-17, 2008.; 2. Cvetana Krstev, Ranka Stanković, Ivan Obradović, Duško Vitas, Miloš Utvić, "Automatic Construction of a Morphological Dictionary of Multi-Word Units", in Proceedings of the 7th International Conference on NLP, IceTAL 2010, Reykjavik, Iceland, August 16-18, 2010. eds. Hrafn Loftsson, Eiríkur Rögnvaldsson, Sigrún Helgadóttir, Lecture Notes in Computer Science 6233 Springer 2010, ISBN 978-3-642-14769-2, pp. 226-237, 2010.
urlDocumentation	korpus.matf.bg.ac.rs/prezentacija

resourceType	lexicalConceptualResource
resourceSubtype	lexicon
mediaType	text
noLanguages	1
multilingualityType	-
languageId	SR
size	96 500
sizeUnit	lemma
annotationType	-

resourceTitle	**Serbian Named Entity Resources**
resourceName	SrpNER
urlDownload	nlp.matf.bg.ac.rs
dateCreation	2005-
projectPartner	UBG
IPRholder.organizationShortName	UBG-MATF
contact.Person.surname	Krstev
contact.Person.givenName	Cvetana
contact.Person.email	cvetana@matf.bg.ac.rs
availability	available-restricted use
license	CC NC BY
resourceLocation	korpus.matf.bg.ac.rs
distributionMedium	DVD-R
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Dusko Vitas
foreseenUse.foreseenUse	NLP applications
foreseenUse.useNLPspecific	various
actualUse.actualUse	NLP applications
actualUse.useNLPspecific	various
description	The Serbian Named Entity resources comprise of dictionaries that contain of approximately 35 000 simple and multi-word unit proper names (lemmas) that allow automatic generation of inflected forms using finite state transducers. All produced word forms are supplied by detailed morpholo-syntactic description. Besides that, all lemmas are equiped with various semantic codes appropriate to named entities. For full and precise recognition of named entities a large collection of finite state transducers is developed.

relevantPublications	1. Gordana Pavlović-Lažetić, Duško Vitas, Cvetana Krstev, “Towards Full Lexical Recognition”, in Proceedings of the 7th International Conference TSD 2004 : Text, Speech and Dialogue, Brno, Czech Republic, September 8-11, 2004, eds. Petr Sojka, Ivan Kopček, Karel Pala, serija “Lecture Notes in Artificial Intelligence” : Subseries of Lecture Notes in Computer Science, eds. J.G. Carbonell, J. Siekmann, pp. 179-186, Springer, Berlin, Heidelberg, 2004.; 2. Cvetana Krstev, Duško Vitas, Ivan Obradović, Miloš Utvić, "E-Dictionaries and Finite-State Automata for the Recognition of Named Entities", FSMNLP, Blois 2011.
urlDocumentation	korpus.matf.bg.ac.rs/prezentacija
resourceType	lexicalConceptualResource
resourceSubtype	lexicon
mediaType	text
noLanguages	1
multilingualityType	-
languageId	SR
size	35 000
sizeUnit	lemma
annotationType	-

resourceTitle	**AlfaNum Morphologic Dictionary of Serbian**
resourceName	AlfaNum MD
urlDownload	alfanum.ftn.uns.ac.rs
dateCreation	2001 – 2010
projectPartner	AlfaNum
IPRholder.organizationShortName	AlfaNum - Speech Technologies, Novi Sad
contact.Person.surname	Sečujski
contact.Person.givenName	Milan
contact.Person.email	secujski@uns.ac.rs
availability	notAvailable
license	-
resourceLocation	alfanum.ftn.uns.ac.rs
distributionMedium	-
restrictionsOfUse	-
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	NLP applications
foreseenUse.useNLPspecific	development of various speech technologies
actualUse.actualUse	NLP applications
actualUse.useNLPspecific	development of various speech technologies

description	This resource is developed for the purpose of linguistic research and development of commercial applications of speech and other language technologies. It consists of 100,517 lexemes (3,888,407 inflected forms)
relevantPublications	1. Milan Sečujski: "Automatic Part-of-Speech Tagging of Texts in Serbian" (PhD Thesis), Faculty of Technical Sciences, Novi Sad, 2009
urlDocumentation	no online documentation provided
resourceType	lexicalConceptualResource
resourceSubtype	lexicon
mediaType	text
noLanguages	1
multilingualityType	-
languageId	SR
size	100517
sizeUnit	lemma
annotationType	-

resourceTitle	**AlfaNum Text Corpus of Serbian**
resourceName	AlfaNumKor
urlDownload	alfanum.ftn.uns.ac.rs
dateCreation	2006 – 2009
projectPartner	AlfaNum
IPRholder.organizationShortName	AlfaNum - Speech Technologies, Novi Sad
contact.Person.surname	Sečujski
contact.Person.givenName	Milan
contact.Person.email	secujski@uns.ac.rs
availability	notAvailable
license	-
resourceLocation	alfanum.ftn.uns.ac.rs
distributionMedium	-
restrictionsOfUse	-
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	NLP applications
foreseenUse.useNLPspecific	development of various speech technologies
actualUse.actualUse	NLP applications
actualUse.useNLPspecific	development of various speech technologies
description	This resource is developed for the purpose of linguistic research and development of commercial applications of speech and other language technologies. It consists of 200,027 words with accentuation and PoS tags.
relevantPublications	1. Milan Sečujski: "Automatic Part-of-Speech Tagging of Texts in Serbian" (PhD Thesis), Faculty of Technical Sciences, Novi Sad, 2009

urlDocumentation	no online documentation provided
resourceType	corpus
resourceSubtype	corpus
mediaType	text
noLanguages	1
multilingualityType	-
languageId	SR
size	200027
sizeUnit	word
annotationType	PoS, accentuation

resourceTitle	**AlfaNum Speech Databases for ASR**
resourceName	AlfaNum ASR
urlDownload	alfanum.ftn.uns.ac.rs
dateCreation	1996 –
projectPartner	AlfaNum
IPRholder.organizationShortName	AlfaNum - Speech Technologies, Novi Sad
contact.Person.surname	Delić
contact.Person.givenName	Vlado
contact.Person.email	vdelic@uns.ac.rs
availability	partially available-unrestricted use; partially notAvailable
license	-
resourceLocation	alfanum.ftn.uns.ac.rs
distributionMedium	-
restrictionsOfUse	partially available-unrestricted use; partially notAvailable
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	NLP applications
foreseenUse.useNLPspecific	automatic speech recognition
actualUse.actualUse	NLP applications
actualUse.useNLPspecific	automatic speech recognition
description	This resource consists of five speech databases: 1. W50S200 contains studio recordings of 200 speakers (53 words per speaker); 2. W150tf1000 contains recordings over the telephone channel (150 words, 600-1000 pronunciations per word); 3. AN_SPEAKER contains office recordings using computer microphone of 44 speakers (29 male, 15 female); 4. S70W100s120 contains recordings originally done in a studio using reel-to-reel tape recorder, later converted to digital format (120 speakers, 70 sentences + 100 isolated utterances per speaker); 5. AlfaNum SpeechDatII(E) contains recordings over the telephone channel (500 speakers, 50 utterances per speaker, according to the SpeechDatII standard). Databases AN_SPEAKER and S70W100s120 are publiclly available, W50S200, W150tf1000 and AlfaNum SpeechDatll(E) are not available.

relevantPublications	1. Vlado Delić: ”Serbian Speech Databases Recorded Within the AlfaNum Project”, Proceedings of DOGS 2000, Novi Sad, 2000, pp. 29-32; 2. Ivan Jokić, Tomislav Dobrijević, Nikša Jakovljević, Vlado Delić: “A Description of a Speech Database intended for Speaker Recognition in Serbian”, Proceedings of TELFOR 2009, Belgrade, 2009.; 3. Nikola Đurić, Darko Pekar, Ljubomir Jovanov: "Structure of the Serbian SpeechDat(E) Database Recorded over Public Telephone Network"; Proceedings of DOGS 2002, Becej, Serbia, 2002, pp. 57-60.
urlDocumentation	no online documentation provided
resourceType	corpus
resourceSubtype	speech corpus
mediaType	audio
noLanguages	1
multilingualityType	-
languageId	SR
size	-
sizeUnit	-
annotationType	phoneme labels, markers of inaccurate pronunciation

resourceTitle	**AlfaNum Speech Databases for TTS**
resourceName	AlfaNum TTS
urlDownload	alfanum.ftn.uns.ac.rs
dateCreation	2000-
projectPartner	AlfaNum
IPRholder.organizationShortName	AlfaNum - Speech Technologies, Novi Sad
contact.Person.surname	Delić
contact.Person.givenName	Vlado
contact.Person.email	vdelic@uns.ac.rs
availability	notAvailable
license	-
resourceLocation	alfanum.ftn.uns.ac.rs
distributionMedium	-
restrictionsOfUse	-
licenseSignatory.Person.position	-
foreseenUse.foreseenUse	NLP applications
foreseenUse.useNLPspecific	text-to-speech synthesis
actualUse.actualUse	NLP applications
actualUse.useNLPspecific	text-to-speech synthesis

description	This resource consists of five speech databases: 1. TTSlab2g2s contains corpus of diphones and dissylables isolated from logatomes and meaningful words recorded in a studio (961 diphone and 625 dissylables); 2. TTSlsMarina, 3. TTSlsSandra, and 4. TTSlsMarija represent corpora composed of a collection of meaningful texts and meaningless word sequences recorded in a studio (each of them comprising of approximately 2 hours of speech); 5. TTSlsSnezana represent corpora composed of a collection of meaningful texts recorded in a studio (Approximately 12 hours of speech).
relevantPublications	1. Milan Secujski, Radovan Obradovic, Darko Pekar, Ljubomir Jovanov, and Vlado Delic: "AlfaNum System for Speech Synthesis in Serbian Language"; Proceedings of TSD 2002, Brno, Czech Republic, 2002, pp. 237-244.; 2. Milan Sečujski, Vlado Delić, Darko Pekar, Radovan Obradović, Dragan Knežević: "An Overview of the AlfaNum Text-to-Speech Synthesis System", Proceedings of SPECOM 2007, Moscow, Russia, 2007, pp. 3-7 (Add. Vol.); 3. Milan Sečujski, Darko Pekar, Dragan Knežević, Vladimir Svrkota, "Prosody prediction in speech synthesis based on regression trees", Proceedings of SinFonIJA 2010, Novi Sad, 2010.
urlDocumentation	no online documentation provided
resourceType	corpus
resourceSubtype	speech corpus
mediaType	audio
noLanguages	1
multilingualityType	-
languageId	SR
size	18
sizeUnit	hour
annotationType	phoneme labels, position of stress, markers of inaccurate pronunciation

resourceTitle	**Digital Archive of the Institute for Balkan Studies**
resourceName	DABI
urlDownload	www.balkaninstitut.com
dateCreation	2007 -
projectPartner	BI-SANU
IPRholder.organizationShortName	BI-SANU
contact.Person.surname	Sikimić
contact.Person.givenName	Biljana
contact.Person.email	Biljana.Sikimic@bi.sanu.ac.rs
availability	available-restricted use
license	according to regulations of the Institute
resourceLocation	www.balkaninstitut.com
distributionMedium	DVD-R
restrictionsOfUse	academic-nonCommercialUse

licenseSignatory.Person.position	Biljana Sikimić
foreseenUse.foreseenUse	human use
foreseenUse.useNLPspecific	various
actualUse.actualUse	human use
actualUse.useNLPspecific	various
description	This database contains digitized audio, video, photo and textual ethnographic material collected in Serbia in the scope of the field research.
relevantPublications	-
urlDocumentation	no online documentation provided
resourceType	corpus
resourceSubtype	multimedia database
mediaType	multimedia
noLanguages	4
multilingualityType	-
languageId	SR; RO; BG; ROM
size	2000
sizeUnit	hour
annotationType	-

resourceTitle	**Corpus of Serbian Language**
resourceName	CSL
urlDownload	www.serbian-corpus.edu.rs/ns/eindex.htm
dateCreation	1996 -
projectPartner	UBG-FF
IPRholder.organizationShortName	UBG-FF
contact.Person.surname	Kostić
contact.Person.givenName	Aleksandar
contact.Person.email	akostic@f.bg.ac.yu
availability	available-restricted use
license	according to regulations of the Laboratory
resourceLocation	www.serbian-corpus.edu.rs/ns/eindex.htm
distributionMedium	-
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Aleksandar Kostić
foreseenUse.foreseenUse	human use
foreseenUse.useNLPspecific	various
actualUse.actualUse	human use
actualUse.useNLPspecific	various

description	The Corpus of Serbian Language CSL was compiled from a sample of 11 million words and spans the Serbian language from the 12th century to the present day. Each word in the CSL is manually tagged for its grammatical status (at the level of inflected morphology), number of graphemes and syllables and phonological structure. The text is also tagged for the beginning and end points of sentences and paragraphs. The system of tagging consists of about 2000 grammatical (inflected) forms.
relevantPublications	1. Ševa Nada, Kostić Aleksandar Đ., Annotated corpus and the empirical evaluation of probability estimates of grammatical forms, Psihologija, 2003, vol. 36, iss. 3, pp. 255-270
urlDocumentation	no online documentation provided
resourceType	corpus
resourceSubtype	corpus
mediaType	text
noLanguages	1
multilingualityType	-
languageId	SR
size	11000000
sizeUnit	word
annotationType	segmentation; PoS; lemmatized

4.6. Appendix 6 - Slovak

resourceTitle	**Slovak National Corpus**
resourceName	prim
urlDownload	http://korpus.juls.savba.sk
dateCreation	ongoing work
projectPartner	LSIL
IPRholder.organizationShortName	LSIL/various
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
availability	available (pseudocorpus)
license	other
resourceLocation	LSIL
distributionMedium	internetBrowsing
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	various
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	various
description	The Slovak National Corpus (SNK) is a representative corpus of contemporary Slovak language written texts. SNK currently contains about 770 million words from broad variety of texts published since 1955 (1953 being the time of most recent substantial Slovak language orthography reform). The corpus is automatically lemmatised and MSD tagged. The documents are annotated with their genre, style and other bibliographic information. There are specialised subcorpora containing fiction, informational texts, professional texts, original Slovak fiction, texts written from 1955 to 1989, and a balanced subcorpus.
relevantPublications	ŠIMKOVÁ, Mária: Slovenský národný korpus – východiská a plány. In: Slovenčina na začiatku 21. storočia. Ed. Mária Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004, p. 150 – 158.; HORÁK, Alexander, GIANITSOVÁ, Lucia, ŠIMKOVÁ, Mária, ŠMOTLÁK, Martin, GARABÍK, Radovan: Slovak National Corpus. In: Text, Speech and Dialogue. 7th International Conference TSD 2004. Proceedings. Ed. P. Sojka, I. Kopeček, K. Pala. Berlin, Heidelberg: Springer – Verlag 2004, p. 89 – 94.; GARABÍK, Radovan: Štruktúra dát v Slovenskom národnom korpuse a ich vonkajšia anotácia. In: Slovenčina na začiatku 21. storočia. Ed. Mária Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004, p. 164 – 173.
urlDocumentation	http://korpus.juls.savba.sk
resourceType	corpus
resourceSubtype	reference corpus
mediaType	text
noLanguages	1
multilingualityType	-

languageId	slk
size	7 700 000
sizeUnit	token
annotationType	segmentation; lemmatization; PosTagging; MSD

resourceTitle	**Corpus of Spoken Slovak**
resourceName	hovor
urlDownload	https://data.juls.savba.sk/oral/
dateCreation	ongoing work
projectPartner	LSIL
IPRholder.organizationShortName	LSIL
contact.Person.surname	Gajdošová
contact.Person.givenName	Katarína
contact.Person.email	katarinag@korpus.juls.savba.sk
availability	available-unrestricted use
license	GNU FDL v1.3, Affero GPL v3, CC-BY-SA
resourceLocation	LSIL
distributionMedium	internetBrowsing
restrictionsOfUse	-
licenseSignatory.Person.position	director
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	various
actualUse.actualUse	human use
actualUse.useNLPspecific	-
description	Corpus of spoken Slovak is a corpus of sound recordings of different types of speech, with the emphasis on spontaneous speech. The recordings are transcribed orthographically and phonemically.
relevantPublications	GARABÍK, Radovan, RUSKO, Milan: Corpus of Spoken Slovak Language. In: Computer Treatment of Slavic and East European Languages. Proceedings of the conference Slovko 2007. Eds. J. Levická, R. Garabík. Brno: Tribun 2007.; GARABÍK, Radovan, KARČOVÁ, Agáta, ŠIMKOVÁ, Mária, GAJDOŠOVÁ, Katarína: Hovorený korpus slovenčiny. In: Čeština v mluveném korpusu. Ed. M. Kopřivová, M. Waclawičová. Praha: Nakladatelství Lidové noviny 2008, p. 227 – 233.
urlDocumentation	https://data.juls.savba.sk/oral/
resourceType	-
resourceSubtype	-
mediaType	-
noLanguages	-
multilingualityType	-
languageId	-
size	-
sizeUnit	-

annotationType	-

resourceTitle	**Slovak Morphological Lexicon**
resourceName	ma
urlDownload	https://data.juls.savba.sk/ma/
dateCreation	-
projectPartner	LSIL
IPRholder.organizationShortName	LSIL/various
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
availability	available-unrestricted use
license	GNU FDL v1.3, Affero GPL v3, CC-BY-SA
resourceLocation	LSIL
distributionMedium	internetBrowsing; download
restrictionsOfUse	attribution; shareAlike
licenseSignatory.Person.position	director
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	morphosyntactic analysis
actualUse.actualUse	human use; NLP applications
actualUse.useNLPspecific	morphosyntactic analysis
description	The lexicon contains full paradigms of 77000 lemmas, together with MSD tags, as used in the Slovak National Corpus. The lexicon serves as a basis for automatic morphological analysis and desambiguation.
relevantPublications	GARABÍK, Radovan: Slovak morphology analyzer based on Levenshtein edit operations. In: Proceedings of the WIKT'06 conference, Bratislava 2006, p. 2 – 5.; Garabík, Radovan: Storing Morphology Information in a Wiki. In: Lexicographic Tools and Techniques. Moscow: IITP RAS 2008, p. 55 – 59.
urlDocumentation	http://korpus.juls.savba.sk/morpho.html
resourceType	lexicalConceptualResource
resourceSubtype	lexicon
mediaType	text
noLanguages	1
multilingualityType	-
languageId	slk
size	77 000
sizeUnit	lemma
annotationType	MSD

resourceTitle	**Slovak Treebank**
resourceName	Slovak Treebank
urlDownload	-
dateCreation	-
projectPartner	LSIL
IPRholder.organizationShortName	LSIL
contact.Person.surname	Gajdošová
contact.Person.givenName	Katarína
contact.Person.email	katarinag@korpus.juls.savba.sk
availability	available-restricted use
license	ANA +NC
resourceLocation	LSIL
distributionMedium	-
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	parsing
actualUse.actualUse	human use
actualUse.useNLPspecific	parsing
description	Slovak language treebank consists of 50000 manually syntactically annotated sentences, using the Prague Dependency Treebank methodology (analytical level). Most of the sentences has been annotated by two independent annotators.
relevantPublications	ŠIMKOVÁ, Mária, GARABÍK, Radovan: Синтаксическая разметка в Словацком национальном корпусе. In: Труды международной конференции Корпусная лингвистика – 2006. Sankt-Petersburg: St. Petersburg University Press 2006, p. 389 – 394.; ŠIMKOVÁ, Mária, GAJDOŠOVÁ, Katarína: Slovenský závislostný korpus. In: Gramatika a korpus 2007. Ed. F. Štícha, M. Fried. Praha: Academia 2008. p. 135 – 141.
urlDocumentation	-
resourceType	lexicalConceptualResource
resourceSubtype	treebank
mediaType	text
noLanguages	1
multilingualityType	-
languageId	slk
size	50 000
sizeUnit	sentence
annotationType	segmentation; lemmatization; PosTagging; MSD; dependency analysis

resourceTitle	**Slovak WordNet**
resourceName	wn
urlDownload	-
dateCreation	-
projectPartner	LSIL
IPRholder.organizationShortName	LSIL
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
availability	available-unrestricted use
license	-BY
resourceLocation	LSIL
distributionMedium	internetBrowsing; download
restrictionsOfUse	-
licenseSignatory.Person.position	director
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	semantic analysis
actualUse.actualUse	human use
actualUse.useNLPspecific	-
description	Slovak WordNet is a a network of lexical-semantic relations, an electronic thesaurus with a structure modelled on that of the Princeton WordNet. The WordNet describes the meaning of a lexical unit of one or more words by placing this unit in a network of links which represent such relations as synonymy, hypernyny, meronymy etc. The Slovak WordNet has been built semi-automatically, using information from bilingual Slovak-English dictionary, and the synsets were then manually proofread. The Slovak synsets are mapped to equivalent English Princeton WordNet semantic equivalents, and contain translation into German, Polish and Lithuanian.
relevantPublications	Radovan Garabík: Slovak National Corpus tools and resources. In: Proceedings of the 5th Workshop on Intelligent and Knowledge oriented Technologies (WIKT 2010). Eds. Laclavík, M., Hluchý, L., November 2010, Bratislava
urlDocumentation	-
resourceType	lexicalConceptualResource
resourceSubtype	wordnet
mediaType	text
noLanguages	4
multilingualityType	-
languageId	slk; pol; deu;lit
size	12 500
sizeUnit	synset
annotationType	-

resourceTitle	**Slovak -English Parallel Corpus**
resourceName	sk-en
urlDownload	-
dateCreation	ongoing work
projectPartner	LSIL
IPRholder.organizationShortName	LSIL/various
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
availability	available (pseudocorpus)
license	other
resourceLocation	LSIL
distributionMedium	internetBrowsing
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	machine translation
actualUse.actualUse	NLP
actualUse.useNLPspecific	machine translation
description	The corpus contains original English fiction texts and their Slovak translations, with automatically aligned sentences.
relevantPublications	-
urlDocumentation	-
resourceType	corpus
resourceSubtype	parallel corpus
mediaType	text
noLanguages	2
multilingualityType	parallel
languageId	slk; eng
size	1 500 000
sizeUnit	sentence
annotationType	segmentation; lemmatization; PosTagging; MSD; alignment

resourceTitle	**Slovak –Czech Parallel Corpus**
resourceName	sk-cs
urlDownload	-
dateCreation	ongoing work
projectPartner	LSIL
IPRholder.organizationShortName	LSIL/various

contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
availability	available (pseudocorpus)
license	other
resourceLocation	LSIL
distributionMedium	internetBrowsing
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	machine translation
actualUse.actualUse	human use; NLP
actualUse.useNLPspecific	machine translation
description	The corpus contains mostly fiction translated between Slovak and Czech (in both direction), with small amount of non-fiction texts and some translations from third language into both Czech and Slovak. The texts are automatically sentence-aligned, with some amount of texts aligned manually.
relevantPublications	-
urlDocumentation	-
resourceType	corpus
resourceSubtype	parallel corpus
mediaType	text
noLanguages	2
multilingualityType	parallel
languageId	slk; ces
size	700 000
sizeUnit	sentence
annotationType	segmentation; lemmatization; PosTagging; MSD; alignment

resourceTitle	**Slovak Web Corpus**
resourceName	sk-web
urlDownload	-
dateCreation	ongoing work
projectPartner	LSIL
IPRholder.organizationShortName	LSIL/various
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
availability	available (pseudocorpus)
license	other

resourceLocation	LSIL
distributionMedium	internetBrowsing
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	various
actualUse.actualUse	human use; NLP
actualUse.useNLPspecific	various
description	Web corpus contains texts downloaded from the .sk domain. The texts are automatically lemmatized and morphologically tagged.
relevantPublications	-
urlDocumentation	-
resourceType	corpus
resourceSubtype	corpus
mediaType	text
noLanguages	1
multilingualityType	-
languageId	slk
size	900000000
sizeUnit	token
annotationType	segmentation; lemmatization; PosTagging; MSD

resourceTitle	**Slovak Legal Texts Corpus**
resourceName	legal
urlDownload	-
dateCreation	2011
projectPartner	LSIL
IPRholder.organizationShortName	LSIL; MS SR
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
availability	available (pseudocorpus)
license	other
resourceLocation	LSIL
distributionMedium	internetBrowsing
restrictionsOfUse	-
licenseSignatory.Person.position	director
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	various

actualUse.actualUse	human use; NLP
actualUse.useNLPspecific	various
description	Corpus of legal texts contains the current (2011) body of Slovak Republic laws. The corpus has been prepared in collaboration with the Ministry of Justice of the Slovak Republic.
relevantPublications	-
urlDocumentation	-
resourceType	corpus
resourceSubtype	corpus
mediaType	text
noLanguages	1
multilingualityType	-
languageId	slk
size	146000000
sizeUnit	token
annotationType	segmentation; lemmatization; PosTagging; MSD

resourceTitle	**Slovak –Russian Parallel Corpus**
resourceName	sk-ru
urlDownload	http://korpus.juls.savba.sk/parus/
dateCreation	2006
projectPartner	LSIL
IPRholder.organizationShortName	LSIL/various
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
availability	available (pseudocorpus)
license	other
resourceLocation	LSIL
distributionMedium	internetBrowsing
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	-
actualUse.actualUse	human use; NLP
actualUse.useNLPspecific	-
description	The corpus contains original Russian fiction texts and their Slovak translations, with automatically aligned sentences.

relevantPublications	Garabík, Radovan, Захаров, Виктор Павлович: Параллельный русско-словацкий корпус. In: Труды международной конференции Корпусная лингвистика. Санкт-Петербург: Издательство С.-Петербургского университета 2006, p. 81 – 87.
urlDocumentation	http://korpus.juls.savba.sk/parus/
resourceType	corpus
resourceSubtype	parallel corpus
mediaType	text
noLanguages	2
multilingualityType	parallel
languageId	slk; rus
size	100 000
sizeUnit	sentence
annotationType	segmentation; lemmatization; PosTagging; MSD; alignment

resourceTitle	**Slovak –French Parallel Corpus**
resourceName	sk-fr
urlDownload	http://korpus.juls.savba.sk/frask/
dateCreation	2007
projectPartner	LSIL
IPRholder.organizationShortName	LSIL/various
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
availability	available (pseudocorpus)
license	other
resourceLocation	LSIL
distributionMedium	internetBrowsing
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	-
actualUse.actualUse	human use; NLP
actualUse.useNLPspecific	-
description	The corpus contains original French fiction texts and their Slovak translations, with automatically aligned sentences.
relevantPublications	VASILIŠINOVÁ, Dorota, GARABÍK, Radovan: Parallel French-Slovak Corpus. In: Computer Treatment of Slavic and East European Languages. Proceedings of the conference Slovko 2007. Eds. J. Levická, R. Garabík. Brno: Tribun 2007.
urlDocumentation	http://korpus.juls.savba.sk/frask/

resourceType	corpus
resourceSubtype	parallel corpus
mediaType	text
noLanguages	2
multilingualityType	parallel
languageId	slk; fra
size	21 000
sizeUnit	sentence
annotationType	segmentation; lemmatization; PosTagging; MSD; alignment