CESAR

Central and South-East European Resources
Project no. 271022

Deliverable D2.3b

Report on resources (actually or potentially) available to the consortium

Version No. 1.3
30/01/2012

Document Information

Deliverable number:	D2.3b
Deliverable title:	Report on resources (actually or potentially) available to the consortium (name)
Due date of deliverable:	30/01/2012
Actual submission date of deliverable:	31/01/2012
Main Author(s):	Svetla Koeva (IBL)
Participants:	Tamas Varadi (HASRIL) Tibor Pinter (HASRIL) Szaszák György (BME-TMIT) Radovam Garabik (LSIL) Maciej Ogrodniczuk (IPIPAN) Adam Przepiórkowski (IPIPAN) Piotr Pezik (ULodz) Marko Tadic (FFZG) Dusko Vitas (UBG) Cvetana Krstev (UBG)
Internal reviewer:	Tamas Varadi (HASRIL)
Workpackage:	2
Workpackage title:	Analysis and selection of language resources
Workpackage leader:	IBL
Dissemination Level:	Public
Version:	1.3
Keywords:	language resources, tools for natural language processing, language technologies

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
1.3	31/01/2012	Final	Tamás Váradi	supervision	
1.2	30/01/2012	draft	Tibor Pinter	editing of the text	
1.1	30/01/2012	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 second six month of the project. A more detailed description of the resources is given in the annex.

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 is specifically focused on the assembly of basic language resources for six Central and South-East European languages, all of them considered, by any comparison, less-resourced: four of them (Hungarian, Polish, Bulgarian, Slovak) 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, Polish, 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
multilingualityType	Whether the corpus is parallel or comparable.	parallel; comparable
languageId	Identifier of the language as defined by ISO 639 that is included in the resource or supported by the tool/service. When more than one 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. 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

2.3. Adaptation to the META-SHARE specifications

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

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

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

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

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

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

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

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

3. Resources identified via CESAR between month sixth and month twelve

The D2.3b 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 extensive enough 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:

- Monolingual (Bulgarian) text corpora:
    - Corpus of Colloquial Bulgarian – consists of transcripts of colloquial speech and amounts up to 534 604 words
    - Diachronic corpus of Bulgarian - Corpus of Medieval and Early Modern Bulgarian texts and manuscripts
- Monolingual (Bulgarian) audio / multimedia corpora:
    - Corpus of Spoken Bulgarian – created in 2011, the corpus amounts up to 605 202 words (312 hours) at present.
- Lexical Conceptual Resources:
    - TREFL – Translation Reference Library - TREFL is a portable, multifunctional database management application for Windows, having the combined characteristics of both a Translation Memory System (bilingual databases, fuzzy matching, concordance, alignment, importing and exporting translation memories, etc.) and those of an Internet/Desktop Search Engine (searching, like with Google search, all these words, this exact phrase, I'm feeling Lucky, etc.), plus some elements of semantic search. It is intended to be used as a simple, versatile, portable, effective and customizable reading, writing and translation aid tool capable of managing very large databases.
- Technology Tools:
    - Bulgarian Spell Checker for Windows - The Bulgarian spell checker WinEst for Microsoft Office detects and marks the incorrectly written words in a text

and suggests the most probable candidates to correct the errors. WinEst offers the entire potential of the contemporary spelling correction: proficiently compiled dictionary, which contains over a million and a half words, and replacement suggestions, which are ordered according to their probability. WinEst is based on the Grammar Dictionary of Bulgarian which contains over 85 000 words. The spell checker exploits logic for detection of performance errors (wrong key pressed, letter swapping, skipped letters or extra letters), competence errors and integrates perfectly into the dictionaries used in Microsoft Office. WinEst uses a fast and effective method for searching and detecting the correct words regardless of the text size. The functionality of the product is realized through the use of minimal acyclic deterministic automata and Levenshtein automata, which allow maximum speed, precision and coverage. A distinctive feature of WinEst is it is easy to install and uninstall, and no System restart is required.

- Bulgarian Spell Checker WinEst - The Bulgarian Spell Checker WinEst is integrated as a web service – both the web service integration and the online spelling checking (as an illustration of the integration) are possible. The Spell Checker is based on the construction of a dictionary in a minimal acyclic deterministic automaton and offers replacement suggestions on the basis of Levenshtein automata. WinEst allows the users to check and correct Bulgarian texts on the Internet. The Spell Checker web service can be used in different blogs, chat forums, online shops, media, and everywhere in the creation of Internet contents, so that it will assist the correct writing of Bulgarian texts.

- Chooser - annotation tool - Chooser is an OS independent multi-functional system for linguistic annotation, adaptable to different annotation schemata. The basic annotation functionalities are: (i) fast and easy-to-perform selection; (ii) run-time access to information for the candidate senses such as definition, frequency, the associated wordnet synsets with all the pertaining info – synonyms, gloss, semantic relations, notes on usage, form, etc.; (iii) identification of MWEs with contiguous and non-contiguous constituents and supplying information for them at run-time. The basic functions are enhanced with flexible text navigation strategies - forward and backward navigation over: (i) all words; (ii) non-annotated words; (iii) all instances of a word; (iv) all instances of a sense. Finally, a flexible search strategy allowing both exact match search according to word form or lemma, and regular expression search is integrated. The tool interface features a fully-fledged visualization of the wordnet synsets for the candidate senses available for a selected LU through coupling with the system for wordnet development and exploration Hydra. A unified wordnet representation in Chooser and Hydra is implemented. Chooser provides multiple-user concurrent access and dynamic real-time update in the knowledge base, so that all changes, such as newly-encoded synsets, literals, relations, are updated in both systems and made available to all the users immediately.

- Hydra - tool for developing wordnets - Hydra is a tool for editing, viewing, searching and validating wordnet. The Hydra API for wordnet processing uses abstract language independent of the data representation, the tool supports a multiple-user concurrent access for editing and browsing arbitrary number of monolingual wordnets, it optimizes data visualization as well as enhances editing, undo/redo functions, etc. The search engine works with the wordnet modal language. The language abstracts the internal data representation and is expressive for the most of the tasks in processing wordnets. Provided that a given wordnet property is definable as a formula in the modal language, the tool determines all the objects in the wordnet structure validating the formula, and hence the property, covering an automatic consistency validation. As a platform-independent system, Hydra has been successfully tested under Linux and Windows.

- Bulgarian Sentence splitter - The sentence splitter marks the sentence boundaries in raw Bulgarian text. The sentence splitter applies regular rules and lexicons. Both - regular rules and lexicons - are manually crafted by an expert. Lists of lexicons (for recognizing abbreviations after which there must be or there might be a capital letter, a number, etc. in the middle of the sentence) are applied before the regular rules. The lexicons are compiled by a separate tool - the Lexicon compiler, as minimal acyclic final state automata which allows an effective processing. Sentence borders are represented as a position and length which allows the incoming text to be kept unchanged as well as an easy integration in different systems for annotation.

- Bulgarian Tokeniser - The Bulgarian tokenizer demarcates strings of letters, numbers, punctuation marks, special symbols, combinations of them and empty symbols. Regular patterns are used to recognize some simple cases of named entities that mean dates, fractions, emails, internet addresses, abbreviations, etc. The tokenizer classifies each recognized token (for example: small cyrillic letters, capital latin letters, etc.). The tokenizer utilizes finite state transducers for token recognition and type matching. The token demarcating and token classifying rules are defined and compiled as finite state transducers with a separate tool - the ParseEst.

- RTComp - Real Time Comparison allows effective management of multilingual databases of numerical speech models and graphical representations for direct visual comparison with the results of the real-time acoustic analysis of the language learners' speech.

- SARP- Speech Analyzer Rapid Plot. Plotting vowels in F2-F1 scatter charts with multiple data sets - The SaRP tool, which is an extension to the programme Speech Analyzer version 3 or later, allows managing databases of oral language samples and creating informative charts in an easy and interactive manner.

  Key features:

  - Computer generated feedback on vowel production by language learners.
  - Designed for automatic or semi-automatic (interactive) retrieving of formant values.
  - Easily creates, saves and opens vowel charts. Fully configurable and easy to use.
  - Support for multiple data sets. Vowel charts comparison by superimposing control charts and user charts.
  - Numerical or visual/graphical editing of the charts and quick-commands: create, move, delete, lock/unlock markers.
  - Calculating and representing graphically the mean values.
  - Integrated library of vocal samples.

resourceName	resourceShortName	resourceLocation	resourceType	size	sizeUnit	LingualityType	Outside the consortium
Corpus of Colloquial Bulgarian	BgSpeech	http://bgspeech.net/bg/resources/razg.html	corpus	534 604	word	monolingual	yes
Diachronic corpus of Bulgarian Language	histdict	http://histdict.uni-sofia.bg	corpus			multilingual	yes
Corpus of Spoken Bulgarian	SpokenBg	-	corpus	605 202	word	monolingual	yes
Bulgarian Spell cheker	WinEst	http://dcl.bas.bg/sites/default/files/webfm/WinEst/winestSetup.exe	technologyToolService	1.5M	word	monolingual	no
Bulgarian Spell Checker Web Service	WebEst	http://dcl.bas.bg/est/index_en.php#tabs-5	technologyToolService	1.5M	word	monolingual	no
Chooser - annotation tool	Chooser	http://dcl.bas.bg	technologyToolService			monolingual	no
Hydra - tool for developing wordnets	Hydra	http://dcl.bas.bg	technologyToolService			multilingual	no
Bulgarian Sentence Splitter		http://dcl.bas.bg	technologyToolService			monolingual	no
Bulgarian Tokenizer		http://dcl.bas.bg	technologyToolService			monolingual	no
TREFL – Translation Reference Library	TREFL	http://web.uni-plovdiv.bg/rousni/index_fr.htm	lexical / conceptual resource; technology tool	1 GB	file	multilingual	yes
SARP- Speech Analyzer Rapid Plot. Plotting vowels in F2-F1 scatter charts with multiple data sets	SARP	http://web.uni-plovdiv.bg/rousni/sarp	technology tool	170 MB	file	multilingual	yes
RTComp - Real Time Comparison	RTComp	http://web.uni-plovdiv.bg/rousni/rtcomp	technology tool	10 MB	file	multilingual	yes

Table 4. Summary of the language resources developed in Bulgaria

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

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

- Monolingual Croatian corpora
  - Croatian Web Corpus (hrWaC) is the largest collected corpus for Croatian so far. It was collected in 2011-06 by crawling the whole .hr internet domain yielding 1.2 billion tokens. The corpus has been lemmatised and MSD-tagged automatically using CroTag system (Agić et al., 2008).
  - Corpus of Narodne novine is a constantly growing collection of texts from the Offical Journal of the Republic of Croatia. A part of this collection is included in the Croatian National Corpus, but the rest is being collected in a separate corpus. Text collecting is done by crawling and additional processing such as boilerplate removal, tokenisation, lemmatisation and MSD-tagging.
  - Croatian Dependency Treebank is a part of the Croatian National Corpus (i.e. Croatian part of the Croatian-English Parallel Corpus, CW2000) where 5000 sentences (ca 100,000 tokens) are manually annotated at the analytical layer following the Prague Dependency Treebank formalism adapted to Croatian. The corpus is currently 4000 sentences in size.
  - Croatian Language Corpus (IHJJ) is a large (ca 70 Mw) text collection that features both, synchronic and diachronic Croatian texts. It is assembled from selected text of Croatian language, covering various functional domains and genres. It includes the literature and other written sources from the period of the beginning of the final shaping of the standardization of Croatian language, i.e. from the second half of the 19th century until present time.
- X-lingual Croatian corpora
  - Croatian Translations of Acquis Communautaire is a corpus of texts that is in the process of compiling thanks to the Translation Department of the Ministry of Foreign and European Affairs. Part of Croatian translations of Acquis has been collected from that source, converted to XML following the JRC Acquis DTD and it will be sentence aligned with English and other official languages of EU. This processing workflow will be kept running at least until the accession of Republic of Croatian to EU on 2013-07-01 since at that time all Acquis will be published in Croatian in the Official Journal of the EU and the whole Acquis will be processed at that point.
  - Croatian-English Parallel Web Corpus is a collection of paraellel Croatian-English texts crawled from .hr domain. This corpus was automatically collected and the parallelity of texts expressed as measure on the scale between 0 and 1. Then the collection of parallel-text candidates is being manually inspected for real parallel texts. The initial crawled corpus has ca 253.000 sentence pairs (ca 8 Mw per language).
- Lexical Conceptual Resources
  - Croatian Wordnet (CroWN) counts 8510 synsets at the moment. These sysets are taken from the BCS1-3 as results of BalkaNet project, that ended in 2004, translated and adapted for Croatian. The synsets are linked with relation of synonymity. From there the language specific approach to CroWN is adopted that may deviate to certain extent from the original PWN. CroWN is still work in progress and it is steadily growing.

- Technology Tools / Services
  - CollTerm is a language independent tool for collocation and term extraction. It is an application that collects collocation and term candidates based on nine different cooccurrence measures for multiword units (i.e. collocations) or distributioal differences from large representative corpus by application of the TF-IDF measurement on singleword units. The language dependent part consists of stop-word list and list of MWU MSD-patterns that can be coded with regular expressions as well. The application will be available under Apache 2.0 license.
  - ccExtractor is the tool for extracting translational candidates from weakly or strongly comparable corpora. It uses approach with modelling contexts and mapping two spaces via initial bilingual lexicons.
  - Croatian Lemmatisaton Web Service is an extension of the existing Croatian Lemmatisation Server that functions only with web-form interface or as tailor-made php scrip call. This extension will feature standard web service protocol that will allow pipeline connections.
  - Croatian NERC Web Service will follow the standard web services protocol to process Croatian texts for NEs. The system has been developed within Intex/NooJ development environment (Bekavac, Tadić 2007) and it will be turned into a web service.

resourceName	**resourceShortName**	**resourceLocation**	**resourceType**	**size**	**sizeUnit**	**LingualityType**	**Outside the consortium**
Croatian Web Corpus	hrWaC	#	corpus	1 186 795 086	token	monolingual	no
Corpus of Narodne novine	NN-corp	http://hnk.ffzg.hr/nn	corpus	15 000 000	token	monolingual	no
Croatian Dependency Treebank	HOBS	http://hobs.ffzg.hr	corpus	4500	sentence	monolingual	no
Croatian Language Corpus	Riznica	http://riznica.ihjj.hr	corpus	70 000 000	token	monolingual	no
Croatian Translations of Acquis	hrAcquis	http://hnk.ffzg.hr/hracquis	corpus	60 000 000	token	multilingual	no
Croatian-English Parallel Web Corpus	hr-enWaC	#	corpus	16 000 000	token	multilingual	no
Croatian Wordnet	CroWN	http://hnk.ffzg.hr/crown	lexicalConceptualResource	8510	synset	monolingual	no

Collocation and Term Extractor	CollTerm	#	technologyTool			multilingual	yes
Comparable Corpora Extractor	ccExtractor	#	technologyTool				yes
Croatian Lemmatisation Web Service	CroLem	http://lt.ffzg.hr/crolem	technoloyToolService				no
Croatian NERC Web Service	CroNERC	http://lt.ffzg.hr/cronerc	technoloyToolService				no

Table 5. Summary of the language resources developed in Croatia

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

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

- Monolingual (Hungarian) corpora:
    - Hun_AudioBook_Egri_csillagok_aligned_text – a Semi automatically selected, time aligned high precision texts to free (librivox) audiobook recordings, created in 2012.
    - Hungarian NER Corpus based on Wikipedia – Excpected to be ready by June 2012 the text of the corpus will be auto-generated from Hungarian Wikipedia articles. It will contain Named Entity (NE) tagging according to the CoNLL standard (Person, Organization, Location and Miscellaneous), and additional morphological and shallow syntactic annotation. The corpus will be the largest ever NE-tagged corpus for Hungarian (ca 1.4 million tokens), which can be used for training and testing NE recognizer applications. Thanks to the standard tagset, the performance of systems trained on the hunNERwiki corpus will be comparable with the performance of other state-of-the-art systems. Besides the obvious advantages of fully automatic building and annotation procedure (reducing the annotation cost), the novelty of the corpus is the application of collaboratively constructed resources (Wikipedia, DBpedia).
    - Hungarian Opinion-Tagged Sentence Bank - a human-annotated resource for researching, evaluating and developing opinion mining systems for Hungarian. The resource consists of several thousand sentences selected from Hungarian online newswire, blogs and social media. Named entities are identified in each sentence with automatic NER tools. 5 independent human annotators are asked to indicate what polarity (opinion) is expressed towards

each entity in each sentence (neutral, positive or negative). Created at the end of 2011, the corpus contains 10.000 annotated sentences at present.

- Hungarian webcorpus - 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.
- Hungarian WSD Corpus - contains 300-500 occurrences of 39 word forms that were selected for the purpose of word sense disambiguation. The Hungarian National Corpus and its Heti Világgazdaság (HVG) subcorpus provided the basis for corpus text selection. Texts were annotated by two independent annotators and differences were disambiguated by a third one.
- Szeged Criminal NE Corpus - contains texts on criminal offences which are annotated for named entities. There are two versions of the corpus: one contains tag-for-tag annotation while the other contains tag-for-meaning annotation. At present, it amounts to 540K items.
- Szeged Treebank FX - annotated for light verb constructions manually. This version contains 6,734 occurrences of 1,215 light verb constructions altogether in 82,099 sentences.

- Bilingual and Multilingual (with Hungarian as one language):
  - Hunglish Corpus - a free sentence-aligned Hungarian-English parallel corpus, which at present amounts to about 4,151,000 sentences. The corpus may be searched through a web-based sentence search service. This service has more than 200,000 visits per month.
  - SzegedParalell - contains texts selected on the basis of grammatical and translational criteria. Sentences representing the grammar of the given language (usually taken from language books) and authentic texts are both included in the parallel corpus, thus, the balance is maintained between artificially constructed and natural language structures. Both paragraph and sentence alignment were checked and corrected manually. Present state: 99K sentence alignment units.
  - SzegedParalellFX - constitutes the basis of the SzegedParalellFX, in which light verb constructions are annotated (14,261 sentence alignment units in size containing 1,100 occurrences of light verb constructions).
- Speech databases:
  - Mindentudás Speech Corpus - An audio collection of public lectures in Hungarian, together with transcriptions. The lectures took place as part of the Mindentudás Egyeteme television series.
- Lexical Conceptual Resources:
  - Hungarian Verb Phrase Constructions – a list of verb phrase constructions (VPC) automatically extracted from the Hungarian National Corpus. VPCs consist of a verb and zero or more noun phrases or prepositional phrases either lexically fixed or lexically free. For example 'to take sg into consideration' has a lexically free direct object and a lexically fixed into-PP. The resource also contains frequency information. At present it consists of 6,200 units.

  - morphdb.hu - Hungarian lexical database and morphological grammar. AT present amounts to 400,000 items.

- Technology tools / services:

  - ProSeg - Automatic Prosodic Segmenter - a phonological phrase aligner for speech sound files. Trained initially for Hungarian, but the design concept ensures that it fits a larger set of languages. A language specific retraining may be necessary when using for other languages. The tool helps the analysis of the prosodic structure and can be used in language and speech tecnology research.

  - Hunalign - a sentence aligner that can use bilingual lexicons as a resource, but in the lack of such lexicon, its automatic lexicon-builder ensures that its precision degrades only marginally.

  - Hungarian Language Processing Tools in NooJ – a morphological dictionary (based on the 60,000 lemmata found in the Concise Dictionary of hungarian Language) and NP-chunker rules. The grammar performing the partial syntactic parsing has been implemented in the NooJ corpus-processing environment, as a set of finite-state transducers. It consists of sequences of rules written by linguists. The tool performs sentence- and clause-segmentation, POS-tagging, NP-recognition, predicate-identification and the identification of the other sentence constituents (eg. adverbials). The input text may be any Hungarian raw text or any xml-text compatible with NooJ, and the output may also be exported in xml-format. NooJ is widely used in Hungarian linguistics and language technology: its usege covers a broad scale of morphological, syntactic, lexical, semantic and psychological content analyses. The core dictionaries and grammars were created in 2011, and are consisting of 10 files at present.

  - Hungarian Phonetic Transcriber - a phonetic transcriber tool using the Hungarian SAMPA character set for the phonetic transcription.

  - Hunmorph – an open source tool and programming library for stemming and morphological analysis.

  - Hunner - a sequential tagger for NLP using Maximum Entropy Learning and Hidden Markov Models. hunner is huntag's instantiation for Named Entity Recognition

  - Hunpars is a syntactic analyzer for Hungarian.

  - Hunpos is an open source reimplementation of TnT, the well known part-of-speech tagger by Thorsten Brants.

  - Huntoken is an open source tool for tokenization and sentence segmentation.

resourceName	resourceShortName	resourceLocation	resourceType	size	sizeUnit	LingualityType	**Outside the consortium**
Hun_AudioBook_Egri_csillagok_aligned_text	Egri_csillagok_aligned_text	not available yet	corpus	-	-	monolingual	no

Hungarian NER Corpus based on Wikipedia	hunNERwiki	BME MOKK	corpus	ca. 1.4 million	token	monolingual	yes
Hungarian Opinion-Tagged Sentence Bank	OpinHuBank	not yet available	corpus	10 000	annotated sentence	monolingual	yes
Hungarian Webcorpus	Hungarian Webcorpus	http://mokk.bme.hu/resources/webcorpus	corpus	589 000 000	token	monolingual	yes
Hungarian WSD Corpus	HuWSD	http://www.inf.u-szeged.hu/rgai/corpus_hunwsd	corpus	300-500x39	text	monolingual	yes
Szeged Criminal NE Corpus	SzegedCriNE	http://www.inf.u-szeged.hu/rgai/corpus_ne	corpus	540K	token	monolingual	yes
Szeged Treebank FX	Szeged Treebank FX	http://www.inf.u-szeged.hu/rgai/mwe	corpus	82K	sentence	monolingual	yes
Hunglish Corpus	Hunglish Corpus	http://mokk.bme.hu/resources/hunglishcorpus	corpus	4 151 000	sentence	bilingual	yes
SzegedParalell	SzegedParalell	http://www.inf.u-szeged.hu/rgai/corpus_paralell	corpus	99K	sentence alignment units	bilingual	yes
SzegedParalellFX	SzegedParalellFX	http://www.inf.u-szeged.hu/rgai/mwe	corpus	14K	sentence alignment units	bilingual	yes
Mindentudás Speech Corpus	Mindentudás Speech Corpus	http://mokk.bme.hu/resources/mindentudas	corpus	200	hour	monolingual	yes
Hungarian Verb Phrase Constructions	HVPC	RIL HAS	lexical / conceptual resource	6 200	unit	monolingual	no
morphdb.hu	morphdb.hu	http://mokk.bme.hu/resources/morphdb-hu	lexical / conceptual resource	400 000	item	monolingual	yes
Automatic Prosodic Segmenter	ProSeg	not available yet	technology tool / service	-	other	multilingual	no
hunalign	hunalign	http://mokk.bme.hu/resources/hunalign	technology tool / service	-	-	bilingual	yes

Hungarian Language Processing Tools in NooJ	Nooj	http://corpus.nytud.hu/nooj	lexical / conceptual resource, technology tool / Service	~10	file	monolingual	no
Hungarian Phonetic Transcriber	HunPhoner	not available yet	technology tool / service	-	other	monolingual	no
hunmorph	hunmorph	http://mokk.bme.hu/resources/hunmorph	technology tool /service	-	-	monolingual	yes
hunner	hunner	http://mokk.bme.hu/resources/huntag	technology tool / service	-	-	monolingual	yes
hunpars	hunpars	http://mokk.bme.hu/resources/hunpars	technology tool /service	-	-	monolingual	yes
hunpos	hunpos	http://mokk.bme.hu/resources/hunpos	technology tool / service	-	-	monolingual	yes
huntoken	huntoken	http://mokk.bme.hu/resources/huntoken	technology tool / service	-	-	monolingual	yes

Table 6. Summary of the language resources developed in Hungary

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

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

- Bilingual Text corpora (with Polish as one languages):
    - PolRosPC - Polish-Russian Parallel Corpus - Developed at the University of Warsaw in 2011, at present the corpus contains ca. 25 million words of both classical literary works and contemporary newspaper and magazine texts aligned at the level of sentences with bibliographic and structural annotation at the level of text units.
- Speech databases:
    - RadioZakŁódź - Polish Radio Żak and Radio Łódź Speech Corpus – A monolingual corpus which at present contains 50 000 words of text and audio, available at http://www.zak.lodz.pl/.
-

- Lexical Conceptual Resources:
    - DOSEC - Dictionary Of Selected English Collocations – Created in 2011, the dictionary contains at present more than 1.6 million potential collocations extracted from the British National Corpus. For each potential collocation a number of association and dispersion measures were computed and recorded in the dictionary along with annotations of part-of –speech patterns in which they were found. The dictionary is available as a logical dump of a relational database and it can be used to complement paradigmatically oriented lexical databases such as WordNet with syntagmatic information about the phraseological potential of word patterns.
    - DoSPiC - Dictionary of Selected Polish Collocations - Created in 2011, at present the dictionary contains more than 2.5 million potential collocations extracted from the National Corpus of Polish. For each potential collocation a number of association and dispersion measures were computed and recorded in the dictionary along with annotations of part-of –speech patterns in which they were found. The dictionary is available as a logical dump of a relational database and it can be used to complement paradigmatically oriented lexical databases such as WordNet with syntagmatic information about the phraseological potential of word patterns.
    - Polish valency dictionary - a new resource to be created by merging existing valency dictionaries (e.g. the dictionary of prof. Świdziński, its extension by Marcin Woliński and related work by Elżbieta Hajnicz) and their further manual development.
    - Składnica - the result of the Polish Ministry of Science and Higher Education research grant (ended in October 2011) on construction of a treebank for Polish using automatic syntactic analysis. The resource is a treebank of Polish constituents created automatically and then manually corrected. At present it consists of 8227 sentences
- Technology Tools:
    - Morfeusz morphological analyzer - a morphological analyzer using lexical data coming from SGJP – the Grammatical Dictionary of Polish by Zygmunt Saloni, Włodzimierz Gruszczyński, Rober Wołosz and Marcin Woliński. Currently its data are being merged with another morphological dictionary – Morfologik to create PoliMorf, which (after manual revision and extension) is intended to become the richest morphological resource for Polish. Morfeusz tool will be recreated after the merging and cleanup process is finished.
    - Morfologik morphological analyzer - a morphological analyzer using lexical data coming from sjp.pl – a crowd-sourced dictionary of Polish used for Internet word games. Currently its data are being merged with another morphological dictionary – Morfeusz SGJP to create PoliMorf, which (after manual revision and extension) is intended to become the richest morphological resource for Polish. Morfologik tool will be recreated after the merging and cleanup process is finished.

resourceName	resourceShortName	resourceLocation	resourceType	size	sizeUnit	Linguality Type	**Outside the consortium**
Polish-Russian Parallel Corpus	PolRosPC	-	corpus	25 000 000	word	bilingual	yes
Polish Radio Żak and Radio Łódź Speech Corpus	RadioZakŁódź	http://www.zak.lodz.pl/, http://www.radiolodz.pl/	corpus	50 000	word	monolingual	yes
Dictionary Of Selected English Collocations	DOSEC	-	lexical / conceptual resource	1 609 152	entry	monolingual	no
Dictionary of Selected Polish Collocations	DoSPiC	-	lexical / conceptual resource	2 500 000	entry	monolingual	no
Polish valency dictionary	Valency dictionary	-	lexical/conceptual resource	-	-	monolingual	no
Składnica	Składnica	http://zil.ipipan.waw.pl/Składnica	lexical/conceptual resource	8227	sentence	monolingual	no
Morfeusz morphological analyzer	Morfeusz	http://sgjp.pl/morfeusz/dopobrania.html	tool	-	-	monolingual	yes
Morfologik morphological analyzer	Morfologik	-	tool	-	-	monolingual	yes

Table 7. Summary of the language resources developed in Poland

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

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

- Monolingual (Serbian) corpora:
  - ASK - Anthology of Serbian Literature – is being developed by the Teaching Faculty, University of Belgrade. Anthology of Serbian Literature project is a project of digitization of the most important works of Serbian literature. This digital library is freely available. The Anthology of Serbian Literature digital library contains more than 130 works of old and new, folk and author literature: from medieval scripts about the lives of Serbian saints, folk poetry and prose, the most important works of Serbian XVIII and XIX century literature, and great literature works of XX century within the public domain,

to the most important works of the Serbian living authors donated for publication in this edition by the authors themselves.

- EbartArchive - Media Archive Ebart – developed by the Ebart Archive, Belgrade. The EbartArchive full-text database contains articles from 27 daily and weekly newspapers, as well as articles from 16 special newspaper supplements and 17 local newspapers published throughout Serbia. Topics covered include Serbian current events, politics, economics, science, culture, and public life. With archives from 2003 to the present, the database contains approximately 4 million fully indexed articles.

- Bilingual and Multilingual (with Serbian as one language):
  - SrpEngSciKor - English-Serbian Corpus of Abstracts of Scientific Projects – was collected by Serbian Ministry of Education and Science. This bilingual corpus contains abstracts in English and Serbian of all project submissions for fundamental and development research that were submitted to the Ministry of Education and Science for the call for proposals in 2010.
  - EngSrpSloFilmKor - English-Slovenian-Serbian Corpus of Film Subtitles – was developed by the NLP group at the Faculty of Mathematics, University of Belgrade. This corpus contains subtitles for 40 movies in English, Serbian and Slovene. Texts are in XML format and all are aligned at the segment level.
- Lexical / Conceptual Resources:
  - Dict-sr - Serbian (Cyrillic and Latin) Hunspell Spellchecking Dictionary – was developed by the NLP group at the Faculty of Mathematics, University of Belgrade. This resource is a part of the Open Office package for Serbian. It was developed by filtering lexica from Serbian part of the Web in 2007. That way forms actually used on Serbian part of the Web were obtained.

Two of these resources are being developed by the Cesar partner (Faculty of Mathematics, University of Belgrade) – Dict-sr and EngSrpSloFilmKor – while three are being developed outside the consortium. Two resources developed by UBG-MATF will be delivered, while the other are still under negotiation.

resourceName	resourceShortName	resourceLocation	resourceType	size	sizeUnit	LingualityType	**Outside the consortium**
Anthology of Serbian Literature	ASK	www.ask.rs	corpus	130	file	monolingual	no
Media Archive Ebart	EbartArchive	http://www.arhiv.rs/	corpus	4 million	article	monolingual	no
English-Serbian Corpus of Abstracts of Scientific Projects	SrpEngSciKor	-	corpus	350 000	word	bilingual	no
English-Slovenian-Serbian Corpus of Film Subtitles	EngSrpSloFilmKor	http://korpus.matf.bg.ac.rs/EngSrpSloFilmKor	corpus	120	file	multilingual	no
Serbian (Cyrillic and Latin) Hunspell Spellchecking Dictionary	Dict-sr	http://wiki.services.openoffice.org/wiki/Dictionaries#Serbian_.28Serbia.2C_Republic_Srpska.29	lexical / conceptual resource	222 000	token	monolingual	no

Table 8. Summary of the language resources developed in Serbia

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

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

- Monolingual (Slovak) text corpora that can be further classified as:
  - Balanced Slovak Corpus - VYV is a balanced corpus with respect to text type. It contains ⅓ fiction, ⅓ informational text, ⅓ professional text (including popular science). The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed. The corpus is planned to be submitted to META-SHARE.
  - Manually Annotated Slovak Corpus - a manually lemmatized and morphosyntactically annotated corpus. It is used as a basis for NLP tools training (primarily POS tagger and lemmatizer). This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed. The Ľ. Štúr Institute of Linguistics provides the ability to train your own tools, by providing access to the computer cluster (on request). The corpus is planned to be submitted to META-SHARE..
  - SNK – Slovak National Corpus The Slovak National Corpus is a representative corpus of contemporary Slovak language written texts published since 1955 (1953 being the time of most recent substantial Slovak language orthography reform). The corpus is automatically lemmatised and MSD tagged. The documents are annotated with their genre, style and other bibliographic information. There are specialised subcorpora containing fiction, informational texts, professional texts, original Slovak fiction, texts written from 1955 to 1989, and a balanced subcorpus. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.

    Changes regarding the Slovak National Corpus since the first report:

    - Slovak National Corpus has been cleaned of incorrectly converted texts

    - additional texts were included in the corpus (increased the size by 36%), up to final size of 770 million tokens

    - new version 5.0 has been released

    - filter used to discard foreign language text fragments has been tuned for better accuracy

    - existing duplicates have been eliminated

    - web interface to the corpus has been provided (using bonito2/NoSketch engine

    The corpus has been submitted to META-SHARE as part of 1st batch of resources.

- SK-WEB – Slovak Web Corpus Slovak Web Corpus contains texts downloaded from the .sk domain. The texts are automaticaly lemmatized and morphologically tagged. The resource has been developed in collaboration of Ľ. Štúr Institute of Linguistics with Masaryk University in Brno, Czech Republic. The first version of the corpus contains 900 million tokens, and new texts are continuously being downloaded from the .sk domain. The new extended version is expected to be released in 2012.
  Changes since the first report:

  - crawler for the .sk domain has been implemented and tested

  -structure of the corpus archive has been designed

  900 million tokens of previously crawled pages have been incorporated into the corpus (collaboration with the Masaryk University, Brno)

  - The corpus is planned to be submitted to META-SHARE as part of 2nd batch of resources.

- Legal – Slovak Legal Texts Corpus contains entire body of law of the Slovak Republic, it has about 146 million tokens, and its foreseen use is mainly in terminology research. The corpus has been prepared in collaboration with the Ministry of Justice of the Slovak Republic. The corpus contains a lot of other language texts (predominately English and Czech), further filtering is necessary. The corpus is planned to be submitted to META-SHARE as part of 2nd batch of resources.

- Bilingual and Multilingual (with Slovak as one language):

  - SK-CS – The Slovak-Czech Parallel Corpus is a corpus of sentence aligned texts, mostly fiction. The Slovak texts are morphologically annotated and disambiguated using the system applied in the Slovak National Corpus, Czech texts are annotated with the morče tagger. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
    Changes since the first report:

    - a dictionary based interface to query the aligned phrase table constructed out of the corpus has been created and published at the Ľ. Štúr Institute's dictionary portal: http://slovniky.korpus.sk/?d=pskcs

    - The corpus has been submitted to META-SHARE as part of 1st batch of resources.

  - SK-EN – The Slovak-English Parallel Corpus The corpus consists of parallel Slovak and English texts (mostly fiction), with automatic lemmatization, morphological analysis (for Slovak), POS tagging (for English). The corpus consists of original English language books and their Slovak translations. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
    Changes since the first report:

    - the corpus has been extended by Slovak → English translated texts (fiction), in addition to the original English → Slovak direction

    - missing bibliography annotation has been added to the corpus entries

    - texts have been cleaned, incorrect alignments have been corrected

- a dictionary based interface to query the aligned phrase table constructed out of the corpus has been created and published at the Ľ. Štúr Institute's dictionary portal: http://slovniky.korpus.sk/?d=psken

The corpus has been submitted to META-SHARE as part of 1st batch of resources.

- SK-FR - Slovak–French Parallel Corpus contains original French fiction texts and their Slovak translations, with automatically aligned sentences. Further development of the corpus (adding new texts, improving and updating linguistic analysis and search interface, standartisation of formats) will be necessary before it can be considered a high quality resource suitable for the CESAR project.
- SK-RU - Slovak–Russian Parallel Corpus contains original Russian fiction texts and their Slovak translations, with automatically aligned sentences. The corpus has been developed in collaboration of Ľ. Štúr Institute of Linguistics and St. Petersburg State University, Russia. Further development of the corpus (adding new texts, improving and updating linguistic analysis and search interface, standartisation of formats) will be necessary before it can be considered a high quality resource suitable for the CESAR project.

- Speech databases:
  - Hovor - Corpus of Spoken Slovak contains audio records of spontaneous and semi-prepared speech from the entire Slovak territory and their text transcripts. Specific characteristics of spoken language are selectively captured in the transcripts, such as irregular structure of an utterance, pronunciation variants, means of speech modulation, and the presence of non-linguistic elements. The Corpus of Spoken Slovak provides material for research and description of the real form of contemporary standard spoken Slovak. This corpus has been released under following licences (multiple licensing): GNU Free Documentation License version 1.3, Affero General Public License version 3, Creative Commons Attribution – ShareAlike 3.0 Unported License.

    Changes since the first report:

    - new recordings have been included in the corpus (increased the size by 140%), up to final size 1.6 Mtokens

    - new version 3.0 has been released

    - transcription rules have been modified to include additional phenomena and to exclude seldom used tags

    - existing transcriptions have been checked for inconsistencies

    The database has been submitted to META-SHARE as part of 1st batch of resources.

- Lexical Conceptual Resources:
  - Dictionary of Slovak Collocations - aimed at the registration and description of selected multiword lexemes and phrasemes as well as typical collocations with restricted collocability. The dictionary provides an overview of the combinatorial behaviour of words, in the first phase the most frequent nouns

extracted from the Slovak National Corpus. Currently, the database contains information about nouns and (as a separate subproject) particles. Description models on the basis of collocational matrices are elaborated also for verbal, adjectival, adverbial and partical collocations. The dictionary has been developed at the Univerzita sv. Cyrila a Metoda in Trnava in collaboration with Ľ. Štúr Institute of Linguistics. Currently, there are ongoing negotiations regarding the licensing and distribution of the dictionary, with the eventual submitting into the META-SHARE.

- Slovak Morphological 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 disambiguation.

  Changes since the first report:

  - the database markup has been extended to indicate substandard variants of the word forms

  - Slovak Morphological Database is a database of lemmas and their inflected wordforms with MSD tags

  - This corpus has been released under the licences: GNU Free Documentation License version 1.3, Affero General Public License version 3, Creative Commons Attribution – ShareAlike 3.0 Unported License.

  The database has been submitted to META-SHARE as part of 1st batch of resources

- Slovak Terminology Database – monoligual database which at present contains 4500 entries

- 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. The treebank is planned to be submitted to META-SHARE as part of 2nd batch of resources.

- WN - Slovak WordNet is 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.

resourceName	resourceShortName	resourceLocation	resourceType	size	sizeUnit	LingualityType	Outside the consortium
Balanced Slovak Corpus	VYV	LSIL	corpus	247 000 000	token	monolingual	no
Dictionary of Slovak Collocations		http://vronk.net/wicol	corpus	250	entry	monolingual	yes
Manually Annotated Slovak Corpus	MAK	LSIL	corpus	1 200 000	token	monolingual	no

Slovak National Corpus	prim	LSIL	corpus	7 700 000	token	monolingual	no
Slovak Web Corpus	sk-web	LSIL	corpus	900 000 000	token	monolingual	yes
Slovak Legal TextsCorpus	legal	LSIL	corpus	146 000 000	token	monolingual	yes
Slovak-Czech Parallel Corpus	sk-cs	LSIL	corpus	730 000	sentence	bilingual	yes
Slovak-English Parallel Corpus	sk-en	LSIL	corpus	1 500 000	sentence	bilingual	no
Slovak-French Parallel Corpus	sk-fr	LSIL	corpus	21 000	sentence	bilingual	yes
Slovak-Russian Parallel Corpus	sk-ru	LSIL	corpus	100 000	sentence	bilingual	yes
Corpus of Spoken Slovak	hovor	LSIS	corpus	178 (audio), 1 643 000 (text)	hour(audio), token (text)	monolingual	yes
Slovak Morphology Database (Lexicon)	ma	LSIL	Lexical / Conceptual Resource	77 000	lemma	monolingual	yes
Slovak Terminology Database	STD	LSIL	Lexical / Conceptual Resource	4 500	entry	monolingual	yes
Slovak Treebank		LSIL	Lexical / Conceptual Resource	50 000	sentence	monolingual	no
Slovak WordNet	wn	LSIL	Lexical / Conceptual Resource	12 500	synset	multilingual	no

Table 9. Summary of the language resources developed in Slovakia

4. Conclusions

During the reported period 73 resources altogether were developed, updated, or contacted. Approximately half of the resources are corpora (33 text and 4 audio or multimedia). Almost 10% of all the resources are lexical/conceptual databases and approx. 17% are technology tools / services. Again almost half of the resources are monolingual (distributed among the different languages), while the rest are bilingual or multilingual (7,3% vs. 8,76% respectively).

Finally, thirty-eight of the resources are identified outside the consortium, which is 52% of the total number of resources.

Resources per Coutry	Total	By Resource type				By Linguality			
		Text Corpora	Audio Corpora	Lexical / Conceptual Database	technology tool / service	Monlingual	Bilingual	Multilngual	Outside the consortium
Bulgaria	12	2	1	1	8	7	0	5	6
Croatia	11	6	0	1	4	7	0	4	2
Hungary	22	10	1	2	9	17	4	1	17
Poland	8	1	1	4	2	7	1	0	4
Serbia	5	4	0	1	0	3	1	1	0
Slovakia	15	10	1	4	0	10	4	1	9
Total	**73**	**33**	**4**	**13**	**23**	**51**	**10**	**12**	**38**

Table 10. Summary of the reported language resources

5. Annex

5.1. Bulgarian language resources detailed specification

resourceName	Corpus of Colloquial Bulgarian
resourceShortName	BgSpeech
downloadLocation	http://bgspeech.net/bg/resources.html
dateCreation	2004
projectPartner	Institute for Bulgarian Language
iprHolder.organizationName	Sofia University, Faculty of Slavic Studies, Department of Bulgarian
contact.Person.surname	Tisheva
contact.Person.givenName	Yovka
contact.Person.email	yovka.tisheva@abv.bg
DistributionInfo	available-unrestricted use
license	
resourceLocation	http://bgspeech.net/bg/resources/razg.html
distributionAccessMedium	webExecutable downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	
foreseenUse	human use

actualUse	human use
description	
relevantPublications	http://bgspeech.net/bg/publications.html
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	Bg
size	534 604
sizeUnit	word

resourceName	Diachronic corpus of Bulgarian Language
resourceShortName	histdict
iprholder.organizationShortName	
contactPerson.surname	Totomanova
contactPerson.givenName	Anna-Maria
contactPerson.email	atotomanova@abv.bg
DistributionInfo	avaiable-restricted use
license	proprietary
distributionAccessMedium	webExecutable
restrictionsOfUse	other
licenseSignatory.Person.position	
ForeseenUse.foreseenUse	human use
ForeseenUse.useNLPspecific	
ActualUse.actualUse	human use;
ActualUse.useNLPspecific	
Description	Corpus od Medieval and Early Modern Bulgarian texts and manuscripts
resourceType	corpus
mediaType	text
lingualityType	multilingual
multilingualityType	It is planned to be parallel
languageId	chu;bul;grc
size	
sizeUnit	
annotationType	orthographicTranscription structuralAnnotation

resourceName	Corpus of Spoken Bulgarian
resourceShortName	SpokenBg
downloadLocation	not available
dateCreation	2011
projectPartner	Institute for Bulgarian Language
iprHolder.organizationName	Ministry of Education, Focus Fondation, Sofia
contact.Person.surname	Tisheva
contact.Person.givenName	Yovka
contact.Person.email	yovka.tisheva@avb.bg

DistributionInfo	notAvailable underNegotiation
license	
resourceLocation	notAvailable
distributionAccessMedium	notAvailable
restrictionsOfUse	notAvailable
licenseSignatory.Person.position	
foreseenUse	human use
actualUse	human use
description	The Corpus of Spoken Bulgarian was created in 2011 and amounts up to 605 202 words (312 hours) at present.
relevantPublications	
resourceType	corpus
mediaType	text / audio /video
lingualityType	monolingual
languageId	Bg
size	312 hours / 605202 words
sizeUnit	hour / word

resourceName	Bulgarian Spell Checker for Windows
resourceShortName	WinEst
downloadLocation	http://dcl.bas.bg/sites/default/files/webfm/WinEst/winestSetup.exe
dateCreation	2011
projectPartner	Institute for Bulgarian language
iprHolder.organizationName	Institute for Bulgarian language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	available-restrictedUse
license	
resourceLocation	http://dcl.bas.bg/est/
distributionAccessMedium	downloadable
restrictionsOfUse	noModifications
licenseSignatory.Person.position	director
foreseenUse	human use
actualUse	human use

description	The Bulgarian spell checker WinEst for Microsoft Office detects and marks the incorrectly written words in a text and suggests the most probable candidates to correct the errors. WinEst offers the entire potential of the contemporary spelling correction: proficiently compiled dictionary, which contains over a million and a half words, and replacement suggestions, which are ordered according to their probability. WinEst is based on the Grammar Dictionary of Bulgarian which contains over 85 000 words. The spell checker exploits logic for detection of performance errors (wrong key pressed, letter swapping, skipped letters or extra letters), competence errors and integrates perfectly into the dictionaries used in Microsoft Office. WinEst uses a fast and effective method for searching and detecting the correct words regardless of the text size. The functionality of the product is realized through the use of minimal acyclic deterministic automata and Levenshtein automata, which allow maximum speed, precision and coverage. A distinctive feature of WinEst is it is easy to install and uninstall, and no System restart is required.
relevantPublications	
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	BG
size	1.5 mega
sizeUnit	words

resourceName	Bulgarian Spell Checker Web Service
resourceShortName	WebEst
downloadLocation	
dateCreation	2011
projectPartner	Institute for Bulgarian language
iprHolder.organizationName	Institute for Bulgarian language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	available-restrictedUse
license	
resourceLocation	http://dcl.bas.bg/est/index_en.php#tabs-5
distributionAccessMedium	webExecutable
restrictionsOfUse	noModifications
licenseSignatory.Person.position	director
foreseenUse	human use
actualUse	human use

description	The Bulgarian Spell Checker WinEst is integrated as a web service – both the web service integration and the online spelling checking (as an illustration of the integration) are possible. The Spell Checker is based on the construction of a dictionary in a minimal acyclic deterministic automaton and offers replacement suggestions on the basis of Levenshtein automata. WinEst allows the users to check and correct Bulgarian texts on the Internet. The Spell Checker web service can be used in different blogs, chat forums, online shops, media, and everywhere in the creation of Internet contents, so that it will assist the correct writing of Bulgarian texts.
relevantPublications	
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	BG
size	1.5 mega
sizeUnit	words

resourceName	***Chooser - annotation tool***
resourceShortName	Chooser
downloadLocation	
dateCreation	2008
projectPartner	Institute for Bulgarian language
iprHolder.organizationName	Institute for Bulgarian language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	available-restrictedUse
license	
resourceLocation	http://dcl.bas.bg
distributionAccessMedium	other
restrictionsOfUse	noModifications
licenseSignatory.Person.position	director
foreseenUse	human use
actualUse	human use

description	Chooser is an OS independent multi-functional system for linguistic annotation, adaptable to different annotation schemata. The basic annotation functionalities are: (i) fast and easy-to-perform selection; (ii) run-time access to information for the candidate senses such as definition, frequency, the associated wordnet synsets with all the pertaining info – synonyms, gloss, semantic relations, notes on usage, form, etc.; (iii) identification of MWEs with contiguous and non-contiguous constituents and supplying information for them at run-time. The basic functions are enhanced with flexible text navigation strategies - forward and backward navigation over: (i) all words; (ii) non-annotated words; (iii) all instances of a word; (iv) all instances of a sense. Finally, a flexible search strategy allowing both exact match search according to word form or lemma, and regular expression search is integrated. The tool interface features a fully-fledged visualization of the wordnet synsets for the candidate senses available for a selected LU through coupling with the system for wordnet development and exploration Hydra. A unified wordnet representation in Chooser and Hydra is implemented. Chooser provides multiple-user concurrent access and dynamic real-time update in the knowledge base, so that all changes, such as newly-encoded synsets, literals, relations, are updated in both systems and made available to all the users immediately.
relevantPublications	Koeva, S., Leseva, S., Tarpomanova, E., Rizov, B., Dimitrova, T., & Kukova, H. (2010). The Bulgarian Sense-Annotated Corpus – Results and Achievements. In M. Tadic, M. Dimitrova-Vulchanova & S. Koeva (Eds.), Proceedings of the FASSBL-7 Conference (pp. 41-48). Zagreb.
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	
size	
sizeUnit	

resourceName	**Hydra - tool for developing wordnets**
resourceShortName	Hydra
downloadLocation	
dateCreation	2008
projectPartner	Institute for Bulgarian language
iprHolder.organizationName	Institute for Bulgarian language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	available-restrictedUse
license	
resourceLocation	http://dcl.bas.bg
distributionAccessMedium	other
restrictionsOfUse	noModifications
licenseSignatory.Person.position	director

foreseenUse	human use
actualUse	human use
description	Hydra is a tool for editing, viewing, searching and validating wordnet. The Hydra API for wordnet processing uses abstract language independent of the data representation, the tool supports a multiple-user concurrent access for editing and browsing arbitrary number of monolingual wordnets, it optimizes data visualization as well as enhances editing, undo/redo functions, etc. The search engine works with the wordnet modal language. The language abstracts the internal data representation and is expressive for the most of the tasks in processing wordnets. Provided that a given wordnet property is definable as a formula in the modal language, the tool determines all the objects in the wordnet structure validating the formula, and hence the property, covering an automatic consistency validation. As a platform-independent system, Hydra has been successfully tested under Linux and Windows.
relevantPublications	S. Koeva, S. Mihov, and T. Tinchev. 2004. Bulgarian wordnet - structure and validation. Romanian J. Of Inf. Sci. And Technology, 7, No. 1-2:61–78. B. Rizov. 2008. Processing Wordnet with Modal Logic. Proceedings of FASSBL 2008: 93-100.
resourceType	technologyToolService
mediaType	text
lingualityType	multilingual
languageId	
size	
sizeUnit	

resourceName	***Bulgarian Sentence Splitter***
resourceShortName	
downloadLocation	
dateCreation	2009
projectPartner	Institute for Bulgarian language
iprHolder.organizationName	Institute for Bulgarian language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	available-restrictedUse
license	
resourceLocation	http://dcl.bas.bg
distributionAccessMedium	other
restrictionsOfUse	noModifications
licenseSignatory.Person.position	director
foreseenUse	NLP applications
actualUse	NLP applications

description	The sentence splitter marks the sentence boundaries in raw Bulgarian text. The sentence splitter applies regular rules and lexicons. Both - regular rules and lexicons - are manually crafted by an expert. Lists of lexicons (for recognizing abbreviations after which there must be or there might be a capital letter, a number, etc. in the middle of the sentence) are applied before the regular rules. The lexicons are compiled by a separate tool - the Lexicon compiler, as minimal acyclic final state automata which allows an effective processing. Sentence borders are represented as a position and length which allows the incoming text to be kept unchanged as well as an easy integration in different systems for annotation.
relevantPublications	
resourceType	technologyToolService
mediaType	text
linguality Type	monolingual
languageId	BG
size	
sizeUnit	

resourceName	***Bulgarian Tokenizer***
resourceShortName	
downloadLocation	
dateCreation	2009
projectPartner	Institute for Bulgarian language
iprHolder.organizationName	Institute for Bulgarian language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	available-restrictedUse
license	
resourceLocation	http://dcl.bas.bg
distributionAccessMedium	other
restrictionsOfUse	noModifications
licenseSignatory.Person.position	director
foreseenUse	NLP applications
actualUse	NLP applications
description	The Bulgarian tokenizer demarcates strings of letters, numbers, punctuation marks, special symbols, combinations of them and empty symbols. Regular patterns are used to recognize some simple cases of named entities that mean dates, fractions, emails, internet addresses, abbreviations, etc. The tokenizer classifies each recognized token (for example: small cyrillic letters, capital latin letters, etc.). The tokenizer utilizes finite state transducers for token recognition and type matching. The token demarcating and token classifying rules are defined and compiled as finite state transducers with a separate tool - the ParseEst.
relevantPublications	
resourceType	technologyToolService

mediaType	text
lingualityType	monolingual
languageId	BG
size	
sizeUnit	

resourceName	TREFL – Translation Reference Library
resourceShortName	TREFL
downloadLocation	http://web.uni-plovdiv.bg/rousni/index_fr.htm
dateCreation	2007
projectPartner	Institute for Bulgarian Language
iprHolder.organizationName	Plovdiv University „Paisii Hilendarski“
contact.Person.surname	Nikolov
contact.Person.givenName	Roussi
contact.Person.email	roussi.nikolov@gmail.com
DistributionInfo	Database management program : available-restricted use Databases : underNegotiation
license	open source
resourceLocation	http://web.uni-plovdiv.bg/rousni/index_fr.htm
distributionAccessMedium	downloadable
restrictionsOfUse	informResourceOwner academic-nonCommercialUse
licenseSignatory.Person.position	Assoc. Prof. Roussi Nikolov, PhD, Head of the Department of Roman and Germanic Studies, Plovdiv University "Paisii Hilendarski"
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	TREFL is a portable, multifunctional database management application for Windows, having the combined characteristics of both a Translation Memory System (bilingual databases, fuzzy matching, concordance, alignment, importing and exporting translation memories, etc.) and those of an Internet/Desktop Search Engine (searching, like with Google search, all these words, this exact phrase, I'm feeling Lucky, etc.), plus some elements of semantic search. It is intended to be used as a simple, versatile, portable, effective and customizable reading, writing and translation aid tool capable of managing very large databases.
relevantPublications	1. Nikolov, R. & Dommergues, J.-Y. (2008) Les modules d'un système d'aide à la traduction en rapport avec la théorie interprétative, Théorie, Littérature Epistémologie, 25, pp 105-123 1. Roussi Nikolov & Malina DITCHEVA, Една програма-помощник за превод, четене и писане, Plovdiv University "Paissii Hilendarski" - Bulgaria, Scientific Works, Vol. 45, Book 1, 2007 – Philology
resourceType	lexical & conceptual resource technology tool
mediaType	text
lingualityType	multilingual
languageId	EN, FR, BG
size	1.00 GB (textual databases + indexes)
sizeUnit	files

resourceName	RTComp - Real Time Comparison
resourceShortName	RTComp
downloadLocation	http://web.uni-plovdiv.bg/rousni/rtcomp
dateCreation	2012
projectPartner	Institute for Bulgarian Language
iprHolder.organizationName	Plovdiv University „Paisii Hilendarski“
contact.Person.surname	Nikolov
contact.Person.givenName	Roussi
contact.Person.email	roussi.nikolov@gmail.com
DistributionInfo	available-unrestricted use
license	Open source
resourceLocation	http://web.uni-plovdiv.bg/rousni/rtcomp
distributionAccessMedium	downloadable
restrictionsOfUse	informResourceOwner
licenseSignatory.Person.position	Assoc. Prof. Roussi Nikolov, PhD, Head of the Department of Roman and Germanic Studies, Plovdiv University "Paisii Hilendarski"
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	RTComp allows effective management of multilingual databases of numerical speech models and graphical representations for direct visual comparison with the results of the real-time acoustic analysis of the language learners' speech.
relevantPublications	-
resourceType	technology tool
mediaType	audio
lingualityType	multilingual
languageId	EN, FR, BG
size	10 MB
sizeUnit	files

resourceName	SARP- Speech Analyzer Rapid Plot. Plotting vowels in F2-F1 scatter charts with multiple data sets
resourceShortName	SARP
downloadLocation	http://web.uni-plovdiv.bg/rousni/sarp
dateCreation	2007
projectPartner	Institute for Bulgarian Language
iprHolder.organizationName	Plovdiv University „Paisii Hilendarski“
contact.Person.surname	Nikolov
contact.Person.givenName	Roussi
contact.Person.email	roussi.nikolov@gmail.com
DistributionInfo	available-restricted use
license	Open source
resourceLocation	http://web.uni-plovdiv.bg/rousni/sarp

distributionAccessMedium	downloadable
restrictionsOfUse	informResourceOwner
licenseSignatory.Person.position	Assoc. Prof. Roussi Nikolov, PhD, Head of the Department of Roman and Germanic Studies, Plovdiv University “Paisii Hilendarski”
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	The SaRP tool, which is an extension to the programme Speech Analyzer version 3 or later, allows managing databases of oral language samples and creating informative charts in an easy and interactive manner. Key features: · Computer generated feedback on vowel production by language learners. · Designed for automatic or semi-automatic (interactive) retrieving of formant values. · Easily creates, saves and opens vowel charts. Fully configurable and easy to use. · Support for multiple data sets. Vowel charts comparison by superimposing control charts and user charts. · Numerical or visual/graphical editing of the charts and quick-commands: create, move, delete, lock/unlock markers. · Calculating and representing graphically the mean values. · Integrated library of vocal samples.
relevantPublications	1. Nikolov, R. & Dommergues & Élise RYST, SaRP : Un outil de représentations graphiques multi-points et multi-séries des formants vocaliques, Plovdiv University "Paissii Hilendarski" - Bulgaria, Scientific Works, Vol. 45, Book 1, 2007 – Philology 2. Nikolov, R. & Nadine HERRY-BENIT, Spécificités méthodologiques de l’analyse des voyelles dans les voix de femmes, Plovdiv University "Paissii Hilendarski" - Bulgaria, Scientific Works, Vol. 46, Book 1, 2008 – Philology 3. Nikolov, R. & Nadine HERRY-BENIT & Anne TORTEL, Positional determination of the quality of schwa in english, Plovdiv University "Paissii Hilendarski" - Bulgaria, Scientific Works, Vol. 47, Book 1, 2009 – Philology
resourceType	technology tool
mediaType	audio
lingualityType	multilingual
languageId	EN, FR, BG
size	170 MB
sizeUnit	files

5.2. Croatian language resources detailed specification

resourceName	Croatian Web Corpus
resourceShortName	hrWaC
downloadLocation	http://www.nljubesic.net/projects/hrWaC.html
dateCreation	2011
projectPartner	FFZG
iprHolder.organizationName	FFZG

contact.Person.surname	Ljubešić
contact.Person.givenName	Nikola
contact.Person.email	nljubesi@ffzg.hr
DistributionInfo	available-restricted
license	CC BY-NC-SA
resourceLocation	http://www.nljubesic.net/projects/hrWaC.html
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse; commercialUse for a fee
licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
description	The Croatian Web Corpus (hrWaC) is the largest collected corpus for Croatian so far. It was collected in 2011-06 by crawling the whole .hr internet domain yielding 1.2 billion tokens. The corpus has been lemmatised and MSD-tagged automatically using CroTag system.
relevantPublications	Ljubešić, N., Erjavec, T. (2011) hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene // Proceedings of the 14th International Conference Text, Speech and Dialogue (TSD2011), Plzeň, Czech Republic, 1-5 September 2011, Lecture Notes in Artificial Intelligence 6836, Springer, Heidelberg, pp 395-402.
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	hrv
size	1,186,795,086
sizeUnit	token

resourceName	Corpus of Narodne novine
resourceShortName	NN-corp
downloadLocation	http://hnk.ffzg.hr/nn
dateCreation	ongoing work
projectPartner	FFZG
iprHolder.organizationName	FFZG
contact.Person.surname	Tadić
contact.Person.givenName	Marko
contact.Person.email	marko.tadic@ffzg.hr
DistributionInfo	available-restricted
license	CC BY-NC-SA
resourceLocation	http://hnk.ffzg.hr/nn
distributionAccessMedium	not yet available for internet access
restrictionsOfUse	academic-nonCommercialUse; commercialUse for a fee
licenseSignatory.Person.position	

foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
description	The Corpus of Narodne novine is a constantly growing collection of texts from the Offical Journal of the Republic of Croatia. A part of this collection is included in the Croatian National Corpus, but the rest is being collected in a separate corpus. Text collecting is done by crawling and additional processing such as boilerplate removal, tokenisation, lemmatisation and MSD-tagging.
relevantPublications	
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	hrv
size	15,000,000
sizeUnit	token

resourceName	Croatian Dependency Treebank
resourceShortName	HOBS
downloadLocation	http://hobs.ffzg.hr
dateCreation	ongoing work
projectPartner	FFZG
iprHolder.organizationName	FFZG
contact.Person.surname	Tadić
contact.Person.givenName	Marko
contact.Person.email	marko.tadic@ffzg.hr
DistributionInfo	available-restricted
license	CC BY-NC-SA
resourceLocation	http://hobs.ffzg.hr
distributionAccessMedium	download
restrictionsOfUse	academic-nonCommercialUse; commercialUse for a fee
licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
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 currently 4,000 sentence in size.
relevantPublications	Tadić, M. (2006) Croatian Dependency Treebank in Multilingual Context. Readings in Multilinguality: Selected papers for young researchers, Bulgarian Academy of Sciences, Sofia, pp. 125-1. / 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

resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	hrv
size	4500
sizeUnit	sentence

resourceName	Croatian Language Corpus
resourceShortName	Riznica
downloadLocation	http://riznica.ihjj.hr
dateCreation	ongoing work
projectPartner	Institute for Croatian Language and Linguistics (IHJJ)
iprHolder.organizationName	Institute for Croatian Language and Linguistics (IHJJ)
contact.Person.surname	Brozović-Rončević
contact.Person.givenName	Dunja
contact.Person.email	dunja@ihjj.hr
DistributionInfo	available for public acces over web interface
license	CC BY-NC-SA
resourceLocation	http://riznica.ihjj.hr/dokumentacija/index.hr.html
distributionAccessMedium	web interface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
description	The Croatian Language Corpus is assembled from selected text of Croatian language, covering various functional domains and genres. It includes literature and other written sources from the period of the beginning of the final shaping of the standardization of Croatian language, i.e. from the second half of the 19th century on.
relevantPublications	Brozović-Rončević, D., Ćavar, D. (2006) Das Korpus der kroatischen Sprache: Hrvatska jezična mrežna riznica, University in Graz, Austria, 2006-06-19.
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	hrv
size	70,000,000
sizeUnit	token

resourceName	Croatian Translations of Acquis Communautaire
resourceShortName	hrAcquis

downloadLocation	http://hnk.ffzg.hr/hracquis
dateCreation	ongoing work
projectPartner	FFZG
iprHolder.organizationName	FFZG
contact.Person.surname	Tadić
contact.Person.givenName	Marko
contact.Person.email	marko.tadic@ffzg.hr
DistributionInfo	available-unrestricted use
license	CC BY-NC-SA
resourceLocation	http://hobs.ffzg.hr
distributionAccessMedium	download
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
description	The Croatian Translations of Acquis Communautaire (hrAcquis) is a corpus of texts that is in the process of compiling thanks to the Translation Department of the Ministry of Foreign and European Affairs. Part of Croatian translations of Acquis has been collected from that source, converted to XML following the JRC Acquis DTD and it will be sentence aligned with English and other official languages of EU. This processing workflow will be kept running at least until the accession of Republic of Croatian to EU on 2013-07-01 since at that time all Acquis will be published in Croatian in the Official Journal of the EU and the whole Acquis in Croatian will be processed at that point.
relevantPublications	Tadić, M. (2003) Jezične tehnologije i hrvatski jezik, Exlibris, Zagreb.
resourceType	corpus
mediaType	text
lingualityType	parallel
languageId	hrv, eng
size	60,000,000
sizeUnit	token

resourceName	Croatian-English Parallel Web Corpus
resourceShortName	hr-enWaC
downloadLocation	http://www.nljubesic.net/projects/hrenWaC.html
dateCreation	2011
projectPartner	FFZG
iprHolder.organizationName	FFZG
contact.Person.surname	Ljubešić
contact.Person.givenName	Nikola
contact.Person.email	nljubesi@ffzg.hr
DistributionInfo	available-unrestricted use
license	CC BY-NC-SA

resourceLocation	http://www.nljubesic.net/projects/hrenWaC.html
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse; commercialUse for a fee
licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
description	The Croatian-English Parallel Web Corpus (hr-enWaC) is a collection of paraellel Croatian-English texts crawled from .hr domain. This corpus was automatically collected and the parallelity of texts expressed as measure on the scale between 0 and 1. Then the collection of parallel-text candidates is being manually inspected for real parallel texts. The initial crawled corpus has ca 253.000 sentence pairs (ca 8 Mw per language).
relevantPublications	
resourceType	corpus
mediaType	text
lingualityType	parallel
languageId	hrv, eng
size	16,000,000
sizeUnit	token

resourceName	Croatian Wordnet
resourceShortName	CroWN
downloadLocation	http://hnk.ffzg.hr/crown
dateCreation	ongoing work
projectPartner	FFZG
iprHolder.organizationName	FFZG
contact.Person.surname	Raffaelli
contact.Person.givenName	Ida
contact.Person.email	ida.raffaelli@ffzg.hr
DistributionInfo	available-unrestricted use
license	CC BY-NC-SA
resourceLocation	http://hnk.ffzg.hr/crown
distributionAccessMedium	download
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
description	The Croatian Wordnet (CroWN) counts 8510 synsets at the moment. These sysets are taken from the BCS1-3 as results of BalkaNet project, that ended in 2004, translated and adapted for Croatian. The synsets are linked with relation of synonymity. From there the language specific approach to CroWN is adopted that may deviate to certain extent from the original PWN. CroWN is still work in progress and it is steadily growing.

relevantPublications	Raffaelli, I., Tadić, M., Bekavac, B., Agić, Ž. (2008) Building Croatian Wordnet, Proceedings of the Global Wordnet Conference, Szeged, Hungary, pp. 349-359.
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	monolingual
languageId	hrv
size	8510
sizeUnit	synset

resourceName	Collocation and Term Extractor
resourceShortName	CollTerm
downloadLocation	http://www.nljubesic.net/projects/CollTerm.html
dateCreation	2011
projectPartner	FFZG
iprHolder.organizationName	Nikola Ljubešić
contact.Person.surname	Ljubešić
contact.Person.givenName	Nikola
contact.Person.email	nljubesi@ffzg.hr
DistributionInfo	available-unrestricted use
license	Apache 2.0
resourceLocation	http://www.nljubesic.net/projects/CollTerm.html
distributionAccessMedium	downloadable
restrictionsOfUse	following the Apache 2.0 licence
licenseSignatory.Person.position	Nikola Ljubešić
foreseenUse	human use, NLP applications
actualUse	human use, NLP applications
description	CollTerm is a language independent tool for collocation and term extraction. It is an application that collects collocation and term candidates based on nine different co occurrence measures for multiword units (i.e. collocations) or distributional differences from large representative corpus by application of the TF-IDF measurement on singleword units. The language dependent part consists of stop-word list and list of MWU MSD-patterns that can be coded with regular expressions as well.
relevantPublications	
resourceType	tool
mediaType	text
lingualityType	language independent with language dependent module(s)
languageId	
size	–
sizeUnit	–

resourceName	Comparable Corpora Extractor
resourceShortName	ccExtractor
downloadLocation	http://www.nljubesic.net/projects/CollTerm.html
dateCreation	2011
projectPartner	FFZG
iprHolder.organizationName	Nikola Ljubešić
contact.Person.surname	Ljubešić
contact.Person.givenName	Nikola
contact.Person.email	nljubesi@ffzg.hr
DistributionInfo	available-unrestricted use
license	Apache 2.0
resourceLocation	http://www.nljubesic.net/projects/ccExtractor.html
distributionAccessMedium	downloadable
restrictionsOfUse	following the Apache 2.0 licence
licenseSignatory.Person.position	Nikola Ljubešić
foreseenUse	human use, NLP applications
actualUse	human use, NLP applications
description	ccExtractor is the tool for extracting translational candidates from weakly or strongly comparable corpora. It uses approach with modelling contexts and mapping two spaces via initial bilingual lexicons.
relevantPublications	
resourceType	tool
mediaType	text
lingualityType	language independent with language dependent module(s)
languageId	
size	–
sizeUnit	–

resourceName	Croatian Lemmatization Web Service
resourceShortName	CroLem
downloadLocation	http://lt.ffzg.hr/crolem
dateCreation	ongoing work
projectPartner	FFZG
iprHolder.organizationName	FFZG
contact.Person.surname	Agić
contact.Person.givenName	Željko
contact.Person.email	zagic@ffzg.hr
DistributionInfo	available-restricted use
license	CC BY-NC-SA
resourceLocation	http://lt.ffzg.hr/crolem
distributionAccessMedium	web service
restrictionsOfUse	academic-nonCommercialUse; commercialUse for a fee

licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
description	The Croatian Lemmatisaton Web Service is an extension of the existing Croatian Lemmatisation Server that functions only with web-form interface or as tailor-made php scrip call. This extension will feature standard web service protocol that will allow pipeline connections and full disambiguation in the tasks of lemmatization and PoS/MSD-tagging of Croatian texts.
relevantPublications	Agić et al. (2008) Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis // Informatica, 32 (2008), 4; 445-451. / 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.
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	hrv
size	
sizeUnit	

resourceName	Croatian NERC Web Service
resourceShortName	CroNERC
downloadLocation	http://lt.ffzg.hr/cronerc
dateCreation	ongoing work
projectPartner	FFZG
iprHolder.organizationName	FFZG
contact.Person.surname	Bekavac
contact.Person.givenName	Božo
contact.Person.email	bbekavac@ffzg.hr
DistributionInfo	available-restricted use
license	CC BY-NC-SA
resourceLocation	http://lt.ffzg.hr/cronerc
distributionAccessMedium	web service
restrictionsOfUse	academic-nonCommercialUse; commercialUse for a fee
licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
description	The Croatian NERC Web Service follows the standard web services protocol to process Croatian texts for NEs. The system has been developed within Intex/NooJ development environment and turned into a web service.

relevantPublications	Bekavac, B., Tadić, M. (2007) Implementation of Croatian NERC System, Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL2007, Prague, pp. 11-18.
resourceType	technologyToolService
mediaType	text
lingualityType	monolingual
languageId	hrv
size	
sizeUnit	

5.3. Hungarian language resources detailed specification

resourceName	Hun_AudioBook_Egri_csillagok_aligned_text
resourceShortName	Egri_csillagok_aligned_text
downloadLocation	if applicable – not yet
dateCreation	2012
projectPartner	BME-TMIT
iprHolder.organizationName	BME-TMIT
contact.Person.surname	Mihajlik
contact.Person.givenName	Peter
contact.Person.email	mihajlik@tmit.bme.hu
DistributionInfo	available-unrestricted use
license	CC-BY
resourceLocation	Not yet available
distributionAccessMedium	downloadable
restrictionsOfUse	attribution
licenseSignatory.Person.position	head of department
foreseenUse	NLP applications
description	Semi automatically selected, time aligned high precision texts to free (librivox) audiobook recordings
relevantPublications	planned
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	HU
Size	
SizeUnit	

resourceName	Hungarian NER Corpus based on Wikipedia
resourceShortName	hunNERwiki
downloadLocation	mokk.bme.hu/resources
dateCreation	June 2012

projectPartner	Research Institute for Linguistics of Hungarian Academy of Sciences
iprHolder.organizationName	Computer and Automation Research Institute of Hungarian Academy of Sciences
contact.Person.surname	Nemeskey
contact.Person.givenName	Dávid Márk
contact.Person.email	nemeskey.david@sztaki.hu
DistributionInfo	available-unrestricted use
license	Creative Commons Attribution-ShareAlike 3.0 License
resourceLocation	BME MOKK
distributionAccessMedium	downloadable
restrictionsOfUse	attribution shareAlike
licenseSignatory.Person.position	Head of Department, Language Technology Research Group
foreseenUse	NLP applications
actualUse	NLP applications
description	The text of the corpus will be auto-generated from Hungarian Wikipedia articles. It will contain Named Entity (NE) tagging according to the CoNLL standard (Person, Organization, Location and Miscellaneous), and additional morphological and shallow syntactic annotation. The corpus will be the largest ever NE-tagged corpus for Hungarian, which can be used for training and testing NE recognizer applications. Thanks to the standard tagset, the performance of systems trained on the hunNERwiki corpus will be comparable with the performance of other state-of-the-art systems. Besides the obvious advantages of fully automatic building and annotation procedure (reducing the annotation cost), the novelty of the corpus is the application of collaboratively constructed resources (Wikipedia, DBpedia).
relevantPublications	in progress
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	hu
size	ca. 1.4 million
sizeUnit	tokens

resourceName	Hungarian Opinion-Tagged Sentence Bank
resourceShortName	OpinHuBank
downloadLocation	-
dateCreation	2011.11.30
projectPartner	HASRIL
iprHolder.organizationName	GeoX Ltd.
contact.Person.surname	Prajczer
contact.Person.givenName	Tamás
contact.Person.email	prajczer@geox.hu
DistributionInfo	available-unrestricted use
license	CC BY 3.0

resourceLocation	not yet available
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse commercialUse attribution
licenseSignatory.Person.position	CEO
foreseenUse	NLP applications
actualUse	NLP applications
description	The OpinHuBank is a human-annotated resource for researching, evaluating and developing opinion mining systems for Hungarian. The resource consists of several thousand sentences selected from Hungarian online newswire, blogs and social media. Named entities are identified in each sentence with automatic NER tools. 5 independent human annotators are asked to indicate what polarity (opinion) is expressed towards each entity in each sentence (neutral, positive or negative).
relevantPublications	
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	HU
size	10.000 annotated sentences
sizeUnit	sentences

resourceName	Hungarian Webcorpus
resourceShortName	Hungarian Webcorpus
downloadLocation	http://mokk.bme.hu/resources/webcorpus
dateCreation	06/06/04
projectPartner	RILHAS
iprHolder.organizationName	Budapest University of Technology
contact.Person.surname	Varga
contact.Person.givenName	Dániel
contact.Person.email	daniel@mokk.bme.hu
DistributionInfo	available-restricted use
license	CC_BY
resourceLocation	http://mokk.bme.hu/resources/webcorpus
distributionAccessMedium	downloadable
restrictionsOfUse	attribution
licenseSignatory.Person.position	assistant researcher
foreseenUse	NLP applications
actualUse	NLP applications
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.

relevantPublications	Creating open language resources for Hungarian. Halácsy Péter, Kornai András, Németh László, Rung András, Szakadát István, Trón Viktor. In Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004), 2004. ftp://ftp.mokk.bme.hu/Hunglish/doc/lrec04szsz.pdf Web-based frequency dictionaries for medium density languages. Kornai, A, Halácsy, P, Nagy, V, Oravecz, Cs, Trón, V, and Varga, D (2006). In: Proceedings of the 2nd International Workshop on Web as Corpus, EACL-06, pages 1--9. http://people.mokk.bme.hu/%7Ekornai/Papers/webcorp.pdf
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	HU
size	589000000
sizeUnit	tokens

resourceName	Hungarian WSD Corpus
resourceShortName	HuWSD
downloadLocation	http://www.inf.u-szeged.hu/rgai/corpus_hunwsd
dateCreation	2007
projectPartner	RILHAS
iprHolder.organizationName	Szeged University
contact.Person.surname	Vincze
contact.Person.givenName	Veronika
contact.Person.email	vinczev@inf.u-szeged.hu
DistributionInfo	available-restricted use
license	NC-NoReD
resourceLocation	http://www.inf.u-szeged.hu/rgai/corpus_hunwsd
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	The Hungarian WSD corpus contains 300-500 occurrences of 39 word forms that were selected for the purpose of word sense disambiguation. The Hungarian National Corpus and its Heti Világgazdaság (HVG) subcorpus provided the basis for corpus text selection. Texts were annotated by two independent annotators and differences were disambiguated by a third one.
relevantPublications	Vincze, Veronika, Szarvas, György, Almási, Attila, Szauter, Dóra, Ormándi, Róbert, Farkas, Richárd, Hatvani, Csaba, Csirik, János: Hungarian Word-sense Disambiguated Corpus. In: Proceedings of 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco.
resourceType	corpus
mediaType	text

lingualityType	monolingual
languageId	HU
size	300-500x39
sizeUnit	texts

resourceName	Szeged Criminal NE Corpus
resourceShortName	SzegedCriNE
downloadLocation	http://www.inf.u-szeged.hu/rgai/corpus_ne
dateCreation	2008
projectPartner	RILHAS
iprHolder.organizationName	Szeged University
contact.Person.surname	Farkas
contact.Person.givenName	Richárd
contact.Person.email	rfarkas@inf.u-szeged.hu
DistributionInfo	available-restricted use
license	NC-NoReD
resourceLocation	http://www.inf.u-szeged.hu/rgai/corpus_ne
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	
foreseenUse	NLP applications
actualUse	NLP applications
description	The corpus contains texts on criminal offences which are annotated for named entites. There are two versions of the corpus: one contains tag-for-tag annotation while the other contains tag-for-meaning annotation.
relevantPublications	
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	HU
size	540K
sizeUnit	tokens

resourceName	Szeged Treebank FX
resourceShortName	Szeged Treebank FX
downloadLocation	http://www.inf.u-szeged.hu/rgai/mwe
dateCreation	2010
projectPartner	RILHAS
iprHolder.organizationName	Szeged University
contact.Person.surname	Vincze
contact.Person.givenName	Veronika
contact.Person.email	vinczev@inf.u-szeged.hu
DistributionInfo	available-restricted use
license	NC-NoReD
resourceLocation	http://www.inf.u-szeged.hu/rgai/mwe

distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	The Szeged Treebank was annotated for light verb constructions manually. This version contains 6734 occurrences of 1215 light verb constructions altogether in 82,099 sentences.
relevantPublications	Vincze, Veronika; Csirik, János 2010: Hungarian Corpus of Light Verb Constructions. In: Proceedings of COLING 2010, Beijing, China, pp. 1110-1118.
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	HU
size	82K
sizeUnit	sentences

resourceName	Hunglish Corpus
resourceShortName	Hunglish Corpus
downloadLocation	http://mokk.bme.hu/resources/hunglishcorpus
dateCreation	11/11/11
projectPartner	RILHAS
iprHolder.organizationName	Budapest University of Technology
contact.Person.surname	Varga
contact.Person.givenName	Dániel
contact.Person.email	daniel@mokk.bme.hu
DistributionInfo	available-restricted use
license	CC_BY
resourceLocation	http://mokk.bme.hu/resources/hunglishcorpus
distributionAccessMedium	downloadable
restrictionsOfUse	attribution
licenseSignatory.Person.position	assistant researcher
foreseenUse	NLP applications
actualUse	NLP applications
description	Hungarian-English parallel corpus automatically aligned at the sentence level.
relevantPublications	Parallel corpora for medium density languages. Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, Viktor Trón. AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4. http://www.ldc.upenn.edu/Catalog/docs/LDC2008T01/ranlp05.pdf
resourceType	corpus
mediaType	text

lingualityType	bilingual
languageId	HU, EN
size	4151000
sizeUnit	sentences

resourceName	SzegedParalell
resourceShortName	SzegedParalell
downloadLocation	http://www.inf.u-szeged.hu/rgai/corpus_paralell
dateCreation	2007
projectPartner	RILHAS
iprHolder.organizationName	Szeged University
contact.Person.surname	Vincze
contact.Person.givenName	Veronika
contact.Person.email	vinczev@inf.u-szeged.hu
DistributionInfo	available-restricted use
license	NC-NoReD
resourceLocation	http://www.inf.u-szeged.hu/rgai/corpus_paralell
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	The English-Hungarian parallel corpus contains texts selected on the basis of grammatical and translational criteria. Sentences representing the grammar of the given language (usually taken from language books) and authentic texts are both included in the parallel corpus, thus, the balance is maintained between artificially constructed and natural language structures. Both paragraph and sentence alignment were checked and corrected manually.
relevantPublications	Krisztina Tóth, Richárd Farkas, András Kocsor: Hybrid algorithm for sentence alignment of Hungarian-English parallel corpora. Acta Cybernetica 18(3):463-478. (2008)
resourceType	corpus
mediaType	text
lingualityType	bilingual
languageId	HU, EN
size	99K
sizeUnit	sentence alignment units

resourceName	SzegedParalellFX
resourceShortName	SzegedParalellFX
downloadLocation	http://www.inf.u-szeged.hu/rgai/mwe
dateCreation	2010

projectPartner	RILHAS
iprHolder.organizationName	Szeged University
contact.Person.surname	Vincze
contact.Person.givenName	Veronika
contact.Person.email	vinczev@inf.u-szeged.hu
DistributionInfo	available-restricted use
license	NC-NoReD
resourceLocation	http://www.inf.u-szeged.hu/rgai/mwe
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	The SzegedParalell corpus constitutes the basis of the SzegedParalellFX, in which light verb constructions are annotated (14,261 sentence alignment units in size containing 1100 occurrences of light verb constructions).
relevantPublications	Veronika Vincze: Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus. Submitted to LREC 2012.
resourceType	corpus
mediaType	text
lingualityType	bilingual
languageId	HU, EN
size	14K
sizeUnit	Sentence alignment units

resourceName	Mindentudás Speech Corpus
resourceShortName	Mindentudás Speech Corpus
downloadLocation	http://mokk.bme.hu/resources/mindentudas
dateCreation	01/05/12
projectPartner	BME-TMIT
iprHolder.organizationName	Budapest University of Technology
contact.Person.surname	Varga
contact.Person.givenName	Dániel
contact.Person.email	daniel@mokk.bme.hu
DistributionInfo	available-restricted use
license	underNegotiation
resourceLocation	http://mokk.bme.hu/resources/mindentudas
distributionAccessMedium	downloadable
restrictionsOfUse	noDerivatives
licenseSignatory.Person.position	assistant researcher
foreseenUse	NLP applications
actualUse	NLP applications

description	An audio collection of public lectures in Hungarian, together with transcriptions. The lectures took place as part of the Mindentudás Egyeteme television series.
relevantPublications	
resourceType	corpus
mediaType	audio, text
lingualityType	monolingual
languageId	HU
size	200
sizeUnit	hours

resourceName	Hungarian Verb Phrase Constructions
resourceShortName	HVPC
downloadLocation	-
dateCreation	2008-2010
projectPartner	RIL HAS
iprHolder.organizationName	RIL HAS
contact.Person.surname	Sass
contact.Person.givenName	Bálint
contact.Person.email	sass.balint@nytud.mta.hu
DistributionInfo	avaiable-restricted use
license	CC BY NC SA
resourceLocation	RIL HAS
distributionAccessMedium	Hard disc
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	research fellow
foreseenUse	NLP applications
actualUse	human use
description	Hungarian Verb Phrase Constructions is a list of verb phrase constructions (VPC) automatically extracted from the Hungarian National Corpus. VPCs consist of a verb and zero or more noun phrases or prepositional phrases either lexically fixed or lexically free. For example 'to take sg into consideration' has a lexically free direct object and a lexically fixed into-PP. The resource also contains frequency information.

relevantPublications	Sass Bálint, Váradi Tamás, Pajzs Júlia, Kiss Margit: Magyar igei szerkezetek - A leggyakoribb vonzatok és szókapcsolatok szótára. [Hungarian Verb Phrase Constructions - a dictionary of frequent complements and collocations.] Tinta, Budapest, 2010. 504 pages. Pajzs, J. and Sass, B: Towards semi-automatic dictionary making. In: Dykstra, A. and Schoonheim, T., (eds): Proceedings of the XIV. EURALEX International Congress, 2010., 453-462. Sass, Bálint and Pajzs, Júlia. FDVC -- Creating a Corpus-driven Frequency Dictionary of Verb Phrase Constructions for Hungarian. In: Sylviane Granger, Magali Paquot (Eds.) eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009. Cahiers du CENTAL 7. Presses universitaires de Louvain, 2010., 263-272. Sass, Bálint.: A Unified Method for Extracting Simple and Multiword Verbs with Valence Information. In: Angelova G. et al. (eds.): Proceedings of RANLP 2009, Borovec, Bulgária, 2009, 399-403. Sass, Bálint.: The Verb Argument Browser. In: Sojka, P., Horák, A., Kopecek, I., Pala, K. (eds.): 11th International Conference on Text, Speech and Dialog, TSD 2008, Brno, Csehország, 2008, Proceedings. Lecture Notes in Computer Science 5246, 187-192.
resourceType	lexical / conceptual resource
mediaType	text
lingualityType	monolingual
languageId	hu
size	6200
sizeUnit	units

resourceName	morphdb.hu
resourceShortName	morphdb.hu
downloadLocation	http://mokk.bme.hu/resources/morphdb-hu
dateCreation	10/01/06
projectPartner	RILHAS
iprHolder.organizationName	Budapest University of Technology
contact.Person.surname	Varga
contact.Person.givenName	Dániel
contact.Person.email	daniel@mokk.bme.hu
DistributionInfo	available-restricted use
license	CC_BY
resourceLocation	http://mokk.bme.hu/resources/morphdb-hu
distributionAccessMedium	downloadable
restrictionsOfUse	attribution
licenseSignatory.Person.position	assistant researcher
foreseenUse	NLP applications

actualUse	NLP applications
description	Hungarian lexical database and morphological grammar.
relevantPublications	Morphdb. hu: Hungarian lexical database and morphological grammar. V. Trón, P. Halácsy, P. Rebrus, A. Rung, P. Vajda, E. Simon. Proceedings of the LREC 2006. http://www.lrec-conf.org/proceedings/lrec2006/pdf/683_pdf.pdf
resourceType	lexical / conceptual resource
mediaType	text
lingualityType	monolingual
languageId	HU
size	400000
sizeUnit	items

resourceName	Automatic Prosodic Segmenter
resourceShortName	ProSeg
downloadLocation	not available yet
dateCreation	2009
projectPartner	BME-TMIT
iprHolder.organizationName	BME-TMIT
contact.Person.surname	Szaszák
contact.Person.givenName	György
contact.Person.email	szaszak@tmit.bme.hu
DistributionInfo	underNegotiation
license	underNegitoation
resourceLocation	not available yet
distributionAccessMedium	downloadable
restrictionsOfUse	informResourceOwner academic-nonCommercialUse
licenseSignatory.Person.position	Head of department
foreseenUse	NLP applications
actualUse	NLP applications
description	Automatic prosodic segmenter is a phonological phrase aligner for speech sound files. Trained initially for Hungarian, but the design concept ensures that it fits a larger set of languages. A language specific retraining may be necessary when using for other languages. The tool helps the analysis of the prosodic structure and can be used in language and speech tecnology research.
relevantPublications	Vicsi K, Szaszák Gy: Using prosody to improve automatic speech recognition. SPEECH COMMUNICATION 52:(5) pp. 413-426. (2010
resourceType	technology tool / service
mediaType	audio
lingualityType	multilingual
languageId	hun
size	
sizeUnit	other

resourceName	hunalign
resourceShortName	hunalign
downloadLocation	http://mokk.bme.hu/resources/hunalign
dateCreation	10/11/11
projectPartner	RILHAS
iprHolder.organizationName	Budapest University of Technology
contact.Person.surname	Varga
contact.Person.givenName	Dániel
contact.Person.email	daniel@mokk.bme.hu
DistributionInfo	available-restricted use
license	LGPL
resourceLocation	http://mokk.bme.hu/resources/hunalign
distributionAccessMedium	downloadable
restrictionsOfUse	shareAlike
licenseSignatory.Person.position	assistant researcher
foreseenUse	NLP applications
actualUse	NLP applications
description	hunalign is a sentence aligner. It can use bilingual lexicons as a resource, but in the lack of such lexicon, its automatic lexicon-builder ensures that its precision degrades only marginally.
relevantPublications	Parallel corpora for medium density languages. Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, Viktor Trón. AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4. http://www.ldc.upenn.edu/Catalog/docs/LDC2008T01/ranlp05.pdf
resourceType	technology tool / service
mediaType	text
lingualityType	bilingual
languageId	–
size	
sizeUnit	

resourceName	Hungarian Language Processing Tools in NooJ
resourceShortName	NooJ
downloadLocation	http://corpus.nytud.hu/nooj
dateCreation	2011
projectPartner	RILHAS
iprHolder.organizationName	RILHAS
contact.Person.surname	Nagy
contact.Person.givenName	Viktor
contact.Person.email	nagyv@nytud.hu
DistributionInfo	available-unrestricted use
license	GPL
resourceLocation	http://corpus.nytud.hu/nooj
distributionAccessMedium	downloadable

restrictionsOfUse	shareAlike
licenseSignatory.Person.position	developer
foreseenUse	NLP applications
actualUse	NLP applications
description	The Hungarian NooJ contains a morphological dictionary (based on the 60 000 lemmata found in the Concise Dictionary of hungarian Language) and NP-chunker rules. The grammar performing the partial syntactic parsing has been implemented in the NooJ corpus-processing environment, as a set of finite-state transducers. It consists of sequences of rules written by linguists. The tool performs sentence- and clause-segmentation, POS-tagging NP-recognition, predicate-identification and the identification of the other sentence constituents (eg. adverbials). The input text may be any Hungarian raw text or any xml-text compatible with NooJ, and the output may also be exported in xml-format. NooJ is widely used in Hungarian linguistics and language technology: its usege covers a broad scale of morphological, syntactic, lexical, semantic and psychological content analyses.
relevantPublications	Kata Gábor 2010. Creating a Shallow-parsed Hungarian Corpus with NooJ. In: T. Váradi-J. Kuti-M. Silberztein Applications of Finite-State Language Processing: Selected Papers from the 2008 International NooJ Conference, Cambridge Scholars Publishing, 67-76,
resourceType	lexical / conceptual resource, technologyToolService
mediaType	text
lingualityType	monolingual
languageId	hu
size	~10
sizeUnit	files

resourceName	Hungarian Phonetic Transcriber
resourceShortName	HunPhoner
downloadLocation	not available yet
dateCreation	2006
projectPartner	BME-TMIT
iprHolder.organizationName	BME-TMIT
contact.Person.surname	Szaszák
contact.Person.givenName	György
contact.Person.email	szaszak@tmit.bme.hu
DistributionInfo	avaiable-restricted use
license	MS NoRedistribution NonCommercial NoDerivatives
resourceLocation	not available yet
distributionAccessMedium	downloadable
restrictionsOfUse	noModifications informResourceOwner onlyMSmembers academic-nonCommercialUse attribution noDerivatives
licenseSignatory.Person.position	Head of Department
foreseenUse	human use NLP applications

actualUse	human use NLP applications
description	Hungarian Phonetic Transcriber is a phonetic transcriber tool using the Hungarian SAMPA character set for the phonetic transcription.
relevantPublications	-
resourceType	technology tool / service
mediaType	text
lingualityType	monolingual
languageId	hun
size	
sizeUnit	other

resourceName	hunmorph
resourceShortName	hunmorph
downloadLocation	http://mokk.bme.hu/resources/hunmorph
dateCreation	03/01/10
projectPartner	RILHAS
iprHolder.organizationName	Budapest University of Technology
contact.Person.surname	Varga
contact.Person.givenName	Dániel
contact.Person.email	daniel@mokk.bme.hu
DistributionInfo	available-restricted use
license	LGPL
resourceLocation	http://mokk.bme.hu/resources/hunmorph
distributionAccessMedium	downloadable
restrictionsOfUse	shareAlike
licenseSignatory.Person.position	assistant researcher
foreseenUse	NLP applications
actualUse	NLP applications
description	hunmorph is an open source tool and programming library for stemming and morphological analysis.
relevantPublications	Hunmorph: open source word analysis. Viktor Trón, András Kornai, György Gyepesi, László Németh, Péter Halácsy, Dániel Varga. Proceedings of the ACM Workshop on Software, 2005. http://www.aclweb.org/anthology-new/W/W05/W05-11.pdf#page=87
resourceType	technology tool / service
mediaType	text
lingualityType	monolingual
languageId	HU
size	
sizeUnit	

resourceName	hunner
resourceShortName	hunner
downloadLocation	http://mokk.bme.hu/resources/huntag

dateCreation	10/11/05
projectPartner	RILHAS
iprHolder.organizationName	Budapest University of Technology
contact.Person.surname	Varga
contact.Person.givenName	Dániel
contact.Person.email	daniel@mokk.bme.hu
DistributionInfo	available-restricted use
license	LGPL
resourceLocation	http://mokk.bme.hu/resources/huntag
distributionAccessMedium	downloadable
restrictionsOfUse	shareAlike
licenseSignatory.Person.position	assistant researcher
foreseenUse	NLP applications
actualUse	NLP applications
description	huntag is a sequential tagger for NLP using Maximum Entropy Learning and Hidden Markov Models. hunner is huntag's instantiation for Named Entity Recognition
relevantPublications	A Hungarian NP-chunker. Gábor Recski, Dániel Varga. The Odd Yearbook, Budapest. 2009.
resourceType	technology tool / service
mediaType	text
lingualityType	monolingual
languageId	HU
size	
sizeUnit	

resourceName	hunpars
resourceShortName	hunpars
downloadLocation	http://mokk.bme.hu/resources/hunpars
dateCreation	03/01/05
projectPartner	RILHAS
iprHolder.organizationName	Budapest University of Technology
contact.Person.surname	Varga
contact.Person.givenName	Dániel
contact.Person.email	daniel@mokk.bme.hu
DistributionInfo	available-restricted use
license	LGPL
resourceLocation	http://mokk.bme.hu/resources/hunpars
distributionAccessMedium	downloadable
restrictionsOfUse	shareAlike
licenseSignatory.Person.position	assistant researcher
foreseenUse	NLP applications

actualUse	NLP applications
description	hunpars is a syntactic analyzer for Hungarian.
relevantPublications	Mondattani elemző alkalmazás. Babarczy A. – Gábor B. – Hamp G. – Kárpáti A. – Rung A. – Szakadát I., 2005, In Alexin Zoltán – Csendes Dóra (szerk.), III. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, 2005, 20–28.
resourceType	technology tool / service
mediaType	text
lingualityType	monolingual
languageId	HU
size	
sizeUnit	

resourceName	hunpos
resourceShortName	hunpos
downloadLocation	http://mokk.bme.hu/resources/hunpos
dateCreation	10/11/05
projectPartner	RILHAS
iprHolder.organizationName	Budapest University of Technology
contact.Person.surname	Varga
contact.Person.givenName	Dániel
contact.Person.email	daniel@mokk.bme.hu
DistributionInfo	available-restricted use
license	LGPL
resourceLocation	http://mokk.bme.hu/resources/hunpos
distributionAccessMedium	downloadable
restrictionsOfUse	shareAlike
licenseSignatory.Person.position	assistant researcher
foreseenUse	NLP applications
actualUse	NLP applications
description	Hunpos is an open source reimplementation of TnT, the well known part-of-speech tagger by Thorsten Brants.
relevantPublications	HunPos: an open source trigram tagger. Halácsy, P. Kornai, A. Oravecz, Cs. ANNUAL MEETING- ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 2007, CONF 45; VOL 2, pages 2-209-2-212. http://acl.ldc.upenn.edu/P/P07/P07-2053.pdf
resourceType	technology tool / service
mediaType	text
lingualityType	monolingual
languageId	
size	
sizeUnit	

resourceName	huntoken
resourceShortName	huntoken
downloadLocation	http://mokk.bme.hu/resources/huntoken
dateCreation	10/11/05
projectPartner	RILHAS
iprHolder.organizationName	Budapest University of Technology
contact.Person.surname	Varga
contact.Person.givenName	Dániel
contact.Person.email	daniel@mokk.bme.hu
DistributionInfo	available-restricted use
license	LGPL
resourceLocation	http://mokk.bme.hu/resources/huntoken
distributionAccessMedium	downloadable
restrictionsOfUse	shareAlike
licenseSignatory.Person.position	assistant researcher
foreseenUse	NLP applications
actualUse	NLP applications
description	huntoken is an open source tool for tokenization and sentence segmentation.
relevantPublications	
resourceType	technology tool / service
mediaType	text
lingualityType	monolingual
languageId	HU, EN
size	
sizeUnit	

5.4. Polish language resources detailed specification

resourceName	Polish-Russian Parallel Corpus
resourceShortName	PolRosPC
downloadLocation	--
dateCreation	2011
projectPartner	Ulodz
iprHolder.organizationName	University of Warsaw
contact.Person.surname	Łaziński
contact.Person.givenName	Marek
contact.Person.email	m.j.lazinski@uw.edu.pl
DistributionInfo	avaiable-restricted use underNegotiation
license	CC-BY-NC
resourceLocation	
distributionAccessMedium	downloadable

restrictionsOfUse	academic-nonCommercialUse attribution
licenseSignatory.Person.position	
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	The Polish-RussianParallel Corpus is being developed at the University of Warsaw. It contains some 25 million words of both classical literary works and contemporary newspaper and magazine texts aligned at the level of sentences with bibliographic and structural annotation at the level of text units.
relevantPublications	
resourceType	corpus
mediaType	text
linguality Type	bilingual
languageId	POL, RUS
size	25 000 000
sizeUnit	words

resourceName	Polish Radio Żak and Radio Łódź Speech Corpus
resourceShortName	RadioZakŁódź
downloadLocation	
dateCreation	
projectPartner	Ulodz
iprHolder.organizationName	Studenckie Radio Żak, Radio Łódź
contact.Person.surname	
contact.Person.givenName	
contact.Person.email	
DistributionInfo	underNegotiation
license	
resourceLocation	http://www.zak.lodz.pl/, http://www.radiolodz.pl/
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	
foreseenUse	human use NLP applications
actualUse	human use
description	
relevantPublications	
resourceType	corpus
mediaType	text /audio
linguality Type	monolingual
languageId	
size	50 000
sizeUnit	words

resourceName	Dictionary Of Selected English Collocations
resourceShortName	DOSEC
downloadLocation	
dateCreation	2011
projectPartner	Ulodz
iprHolder.organizationName	University of Łódź
contact.Person.surname	Pęzik
contact.Person.givenName	Piotr
contact.Person.email	piotr.pezik@gmail.com
DistributionInfo	avaiable-restricted use
license	CC-BY-NC
resourceLocation	
distributionAccessMedium	accessibleThroughInterface downloadable
restrictionsOfUse	academic-nonCommercialUse attribution
licenseSignatory.Person.position	
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	The Dictionary of Selected English Collocations contains more than 1.6 million potential collocations extracted from the British National Corpus. For each potential collocation a number of association and dispersion measures were computed and recorded in the dictionary along with annotations of part-of –speech patterns in which they were found. The dictionary is available as a logical dump of a relational database and it can be used to complement paradigmatically oriented lexical databases such as WordNet with syntagmatic information about the phraseological potential of word patterns.
relevantPublications	
resourceType	lexical / conceptual resource
mediaType	text
lingualityType	monolingual
languageId	eng
size	1 609 152
sizeUnit	entries

resourceName	Dictionary of Selected Polish Collocations
resourceShortName	DoSPiC
downloadLocation	
dateCreation	2011
projectPartner	Ulodz
iprHolder.organizationName	University of Łódź
contact.Person.surname	Pęzik

contact.Person.givenName	Piotr
contact.Person.email	piotr.pezik@gmail.com
DistributionInfo	avaiable-restricted use
license	CC-BY-NC
resourceLocation	
distributionAccessMedium	accessibleThroughInterface downloadable
restrictionsOfUse	academic-nonCommercialUse attribution
licenseSignatory.Person.position	
foreseenUse	human use /NLP applications
actualUse	human use / NLP applications
description	The Dictionary of Selected Polish Collocations contains more than 2.5 million potential collocations extracted from the National Corpus of Polish. For each potential collocation a number of association and dispersion measures were computed and recorded in the dictionary along with annotations of part-of – speech patterns in which they were found. The dictionary is available as a logical dump of a relational database and it can be used to complement paradigmatically oriented lexical databases such as WordNet with syntagmatic information about the phraseological potential of word patterns.
relevantPublications	
resourceType	lexical / conceptual resource
mediaType	text
lingualityType	monolingual
languageId	POL
size	2 500 000
sizeUnit	entries

resourceName	Polish Valency Dictionary
resourceShortName	Valency dictionary
downloadLocation	–
dateCreation	–
projectPartner	IPIPAN
iprHolder.organizationName	Institute of Computer Science, Polish Academy of Sciences
contact.Person.surname	Przepiórkowski
contact.Person.givenName	Adam
contact.Person.email	adam.przepiorkowski@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	under negotiation, FreeBSD expected
resourceLocation	–
distributionAccessMedium	planned to be downloadable
restrictionsOfUse	under negotiation
licenseSignatory.Person.position	Head of the Linguistic Engineering Group, IPIPAN
foreseenUse	human use, NLP applications

actualUse	NLP applications
description	The valency dictionary will be a new resource created by merging existing valency dictionaries (e.g. the dictionary of prof. Świdziński, its extension by Marcin Woliński and related work by Elżbieta Hajnicz) and their further manual development.
relevantPublications	Elżbieta Hajnicz. Grouping alternating schemata in semantic valence dictionary of Polish verbs. In: Text, Speech and Dialogue: 14th International Conference, TSD 2011, Brno, Czech Republic, volume 6836 of Lecture Notes in Artificial Intelligence, pages 155–162, Heidelberg, 2011. Springer-Verlag.
resourceType	lexical/conceptual resource
mediaType	text
lingualityType	monolingual
languageId	POL
size	unknown yet
sizeUnit	–

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

languageId	POL
size	8227
sizeUnit	sentences

resourceName	Morfeusz Morphological Analyzer
resourceShortName	Morfeusz
downloadLocation	–
dateCreation	1990s (current version)
projectPartner	IPIPAN
iprHolder.organizationName	Marcin Woliński
contact.Person.surname	Woliński
contact.Person.givenName	Marcin
contact.Person.email	marcin.wolinski@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	under negotiation, FreeBSD expected
resourceLocation	http://sgjp.pl/morfeusz/dopobrania.html (current version)
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution
licenseSignatory.Person.position	Head of the Linguistic Engineering Group
foreseenUse	human use, NLP applications
actualUse	human use, NLP applications (current version)
description	Morfeusz is a morphological analyzer using lexical data coming from SGJP – the Grammatical Dictionary of Polish by Zygmunt Saloni, Włodzimierz Gruszczyński, Rober Wołosz and Marcin Woliński. Currently its data are being merged with another morphological dictionary – Morfologik to create PoliMorf, which (after manual revision and extension) is intended to become the richest morphological resource for Polish. Morfeusz tool will be recreated after the merging and cleanup process is finished.
relevantPublications	Marcin Woliński. Morfeusz — a practical tool for the morphological analysis of Polish. In: Mieczysław A. Kłopotek, Sławomir T. Wierzchoń and Krzysztof Trojanowski, editors, Intelligent Information Processing and Web Mining, Advances in Soft Computing, pages 511–520. Springer-Verlag, Berlin, 2006.
resourceType	tool
mediaType	text
lingualityType	monolingual
languageId	POL
size	–
sizeUnit	–

resourceName	Morfologik Morphological Analyzer
resourceShortName	Morfologik
downloadLocation	http://morfologik.blogspot.com (current version)
dateCreation	1990s (current version)
projectPartner	IPIPAN
iprHolder.organizationName	Marcin Miłkowski
contact.Person.surname	Miłkowski
contact.Person.givenName	Marcin
contact.Person.email	marcin.milkowski@ifispan.waw.pl
DistributionInfo	available, unrestricted use
license	under negotiation, FreeBSD expected
resourceLocation	–
distributionAccessMedium	downloadable
restrictionsOfUse	attribution
licenseSignatory.Person.position	–
foreseenUse	human use, NLP applications
actualUse	NLP applications
description	Morfologik is a morphological analyzer using lexical data coming from sjp.pl – a crowd-sourced dictionary of Polish used for Internet word games. Currently its data are being merged with another morphological dictionary – Morfeusz SGJP to create PoliMorf, which (after manual revision and extension) is intended to become the richest morphological resource for Polish. Morfologik tool will be recreated after the merging and cleanup process is finished.
relevantPublications	Marcin Miłkowski. Developing an open-source, rule-based proofreading tool. Software: Practice and Experience, 40(7):543–566, 2010.
resourceType	tool
mediaType	text
lingualityType	monolingual
languageId	POL
size	–
sizeUnit	–

5.5. Serbian language resources detailed specification

resourceName	Anthology of Serbian Literature
resourceShortName	ASK
downloadLocation	www.ask.rs
dateCreation	2009
projectPartner	UBG-UF
iprHolder.organizationName	University of Belgrade, Teacher Faculty
contact.Person.surname	Jovanović

contact.Person.givenName	Aleksandar
contact.Person.email	Aleksandar.jovanovic@uf.bg.ac.rs
DistributionInfo	available-unrestricted use
license	-
resourceLocation	www.ask.rs
distributionAccessMedium	downloadable
restrictionsOfUse	-
licenseSignatory.Person.position	-
foreseenUse	NLP applications
actualUse	human use
description	Anthology of Serbian Literature project is a project of digitization of the most important works of Serbian literature. This digital library is freely available. The Anthology of Serbian Literature digital library contains more than 130 works of old and new, folk and author literature: from medieval scripts about the lives of Serbian saints, folk poetry and prose, the most important works of Serbian XVIII and XIX century literature, and great literature works of XX century within the public domain, to the most important works of the Serbian living authors donated for publication in this edition by the authors themselves.
relevantPublications	-
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	SR
size	130
sizeUnit	files

resourceName	Media Archive Ebart
resourceShortName	EbartArchive
downloadLocation	http://www.arhiv.rs/
dateCreation	2003-
projectPartner	Ebart
iprHolder.organizationName	Ebart - Belgrade
contact.Person.surname	Ćurguz
contact.Person.givenName	Kazimir
contact.Person.email	office@archive.rs
DistributionInfo	underNegotiation
license	-
resourceLocation	http://www.arhiv.rs/
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	-
licenseSignatory.Person.position	-
foreseenUse	NLP applications
actualUse	Human use

description	The EbartArchive full-text database contains articles from 27 daily and weekly newspapers, as well as articles from 16 special newspaper supplements and 17 local newspapers published throughout Serbia. Topics covered include Serbian current events, politics, economics, science, culture, and public life. With archives from 2003 to the present, the database contains approximately 4 million fully indexed articles.
relevantPublications	-
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	SR
size	4 million
sizeUnit	articles

resourceName	English-Serbian Corpus of Abstracts of Scientific Projects
resourceShortName	SrpEngSciKor
downloadLocation	-
dateCreation	July 2010
projectPartner	MON
iprHolder.organizationName	Serbian Ministry of Education and Science
contact.Person.surname	Grubin
contact.Person.givenName	Jasmina
contact.Person.email	jasmina.grubin@nauka.gov.rs
DistributionInfo	underNegotiation
license	-
resourceLocation	-
distributionAccessMedium	-
restrictionsOfUse	-
licenseSignatory.Person.position	-
foreseenUse	NLP applications
actualUse	human use
description	This bilingual corpus contains abstracts in English and Serbian of all project submissions for fundamental and development research that were submitted to the Ministry of Education and Science for the call for proposals in 2010.
relevantPublications	-
resourceType	corpus
mediaType	Text
lingualityType	bilingual
languageId	EN; SR
size	350,000
sizeUnit	words

resourceName	English-Slovenian-Serbian Corpus of Film Subtitles
resourceShortName	EngSrpSloFilmKor
downloadLocation	http://korpus.matf.bg.ac.rs/EngSrpSloFilmKor
dateCreation	2006
projectPartner	UBG-MATF
iprHolder.organizationName	University of Belgrade, Faculty of Mathematics
contact.Person.surname	Vitas
contact.Person.givenName	Duško
contact.Person.email	vitas@matf.bg.ac.rs
DistributionInfo	avaiable-restricted use
license	CC_BY-NC
resourceLocation	http://korpus.matf.bg.ac.rs/EngSrpSloFilmKor
distributionAccessMedium	downloadable
restrictionsOfUse	informResourceOwner academic-nonCommercialUse attribution
licenseSignatory.Person.position	-
foreseenUse	NLP applications
actualUse	NLP applications
description	This corpus contains subtitles for 40 movies in English, Serbian and Slovene. Texts are in XML format and all are aligned at the segment level.
relevantPublications	-
resourceType	Corpus
mediaType	Text
lingualityType	Multilingual
languageId	EN; SR; SI
size	120
sizeUnit	files

resourceName	Serbian (Cyrillic and Latin) Hunspell Spellchecking Dictionary
resourceShortName	Dict-sr
downloadLocation	http://wiki.services.openoffice.org/
dateCreation	2010-08-18
projectPartner	UMG-MATF
iprHolder.organizationName	University of Belgrade
contact.Person.surname	Rakić
contact.Person.givenName	Goran
contact.Person.email	grakic@devbase.net
DistributionInfo	available-unrestricted use
license	disjunctive tri-licence GNU LGPL version 2.1 or later / MPL version 1.1 or later / GNU GPL version 2 or later
resourceLocation	http://wiki.services.openoffice.org/wiki/Dictionaries#Serbian_.28Serbia.2C_Republic_Srpska.29
distributionAccessMedium	downloadable
restrictionsOfUse	attribution shareAlike

licenseSignatory.Person.position	-
foreseenUse	NLP applications
actualUse	human use NLP applications
description	This resource is a part of the Open Office package for Serbian. It was developed by filtering lexica from Serbian part of the Web in 2007. That way forms actually used on Serbian part of the Web were obtained.
relevantPublications	-
resourceType	lexical / conceptual resource
mediaType	Text
lingualityType	Monolingual
languageId	SR
size	222,000
sizeUnit	tokens

5.6. Slovak language resources detailed specification

resourceName	Balanced Slovak Corpus
resourceShortName	VYV
downloadLocation	
dateCreation	2010
projectPartner	LSIL
iprHolder.organizationName	LSIL
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available-restricted use
license	other
resourceLocation	LSIL
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	VYV is a balanced corpus with respect to text type. It contains ⅓ fiction, ⅓ informational text, ⅓ professional text (including popular science). The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
relevantPublications	Radovan Garabík: 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.
resourceType	corpus

mediaType	text
lingualityType	monolingual
languageId	sk
size	247 000 000
sizeUnit	token

resourceName	Dictionary of Slovak Collocations
resourceShortName	
downloadLocation	http://vronk.net/wicol
dateCreation	2010
projectPartner	LSIL
iprHolder.organizationName	LSIL Univerzita sv. Cyrila a Metoda v Trnave, Trnava
contact.Person.surname	Ďurčo
contact.Person.givenName	Peter
contact.Person.email	durco@vronk.net
DistributionInfo	Under negotiation
license	
resourceLocation	http://vronk.net/wicol
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	
licenseSignatory.Person.position	director
foreseenUse	human use
actualUse	human use
description	The dictionary is aimed at the registration and description of selected multiword lexemes and phrasemes as well as typical collocations with restricted collocability. The dictionary provides an overview of the combinatorial behaviour of words, in the first phase the most frequent nouns extracted from the Slovak National Corpus. Currently, the database contains information about nouns and (as a separate subproject) particles. Description models on the basis of collocational matrices are elaborated also for verbal, adjectival, adverbial and partical collocations.
relevantPublications	Peter Ďurčo, Radovan Garabík, Daniela Majchráková, Matej Ďurčo: Dictionary of Slovak Collocations. In: Representing Semantics in Digital Lexicography. Innovative Solutions for Lexical Entry Content in Slavic Lexicography. Warsaw: Institute of Slavic Studies, Polish Academy of Sciences 2009, p. 128 – 137.
resourceType	
mediaType	
lingualityType	monolingual
languageId	sk
size	250
sizeUnit	entries

resourceName	Manually Annotated Slovak Corpus
resourceShortName	MAK
downloadLocation	
dateCreation	2005
projectPartner	LSIL
iprHolder.organizationName	LSIL
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available-restricted use
license	other
resourceLocation	LSIL
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	MAK is a manually lemmatized and morphosyntactically annotated corpus. It is used as a basis for NLP tools training (primarily POS tagger and lemmatizer). This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed. The organization provides the ability to train your own tools, by providing access to the computer cluster (on request).
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, ISBN 978-80-970145-2-0, p. 2 – 7.
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	sk
size	1 200 000
sizeUnit	token

resourceName	Slovak National Corpus
resourceShortName	prim
downloadLocation	http://korpus.juls.savba.sk/
dateCreation	ongoing work
projectPartner	LSIL
iprHolder.organizationName	LSIL
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available-restricted use
license	other

resourceLocation	LSIL
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	human use / NLP applications
actualUse	human use / NLP applications
description	The Slovak National Corpus is a representative corpus of contemporary Slovak language written texts published since 1955 (1953 being the time of most recent substantial Slovak language orthography reform). The corpus is automatically lemmatised and MSD tagged. The documents are annotated with their genre, style and other bibliographic information. There are specialised subcorpora containing fiction, informational texts, professional texts, original Slovak fiction, texts written from 1955 to 1989, and a balanced subcorpus. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
relevantPublications	Mária ŠIMKOVÁ: Slovenský národný korpus ako pomôcka pri výučbe slovenského jazyka. In: K problematike vyučovania materinského jazyka a literatúry II. Ed. M. Vojtech. Bratislava: Univerzita Komenského 2007.; Radovan GARABÍK: Словацкий национальный корпус. In: Труды международной конференции Корпусная лингвистика. Sankt-Petersburg, Russia: St. Petersburg University Press 2004.; R. Garabík, L. Gianitsová, A. Horák, M. Šimková., M. Šmotlák.: Slovak National Corpus. In: Proceedings of the conference TSD 2004. Brno, Czech Republic: Springer-Verlag 2004.; Radovan GARABÍK: 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, s. 164 – 173.; Mária ŠIMKOVÁ: Slovenský národný korpus ako pomôcka pri výučbe slovenského jazyka. In: K problematike vyučovania materinského jazyka a literatúry II. Ed. M. Vojtech. Bratislava: Univerzita Komenského 2007.; Mária ŠIMKOVÁ: Slovak National Corpus – history and current situation. In: Insight into Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2005, s. 152 – 159.; A. HORÁK, L.GIANITSOVÁ, M. ŠIMKOVÁ, M. ŠMOTLÁK, R. GARABÍK: 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, s. 89 – 94.; Mária ŠIMKOVÁ: Čo je možné dozvedieť sa zo Slovenského národného korpusu. In: Čeština doma a ve světě, 2004, roč. 12, č. 3 – 4, s. 130 – 145.; Mária ŠIMKOVÁ: Možnosti využitia Slovenského národného korpusu na štúdium slovenského jazyka. In: Studia Academica Slovaca 33. Prednášky z XL. letnej školy slovenského jazyka a kultúry. Ed.: Jozef Mlacek – Miloslav Vojtech. Bratislava: Filozofická fakulta Univerzity Komenského 2004, s. 204 – 218.; Mária ŠIMKOVÁ: Slovenský národný korpus – východiská a plány. In: Slovenčina na začiatku 21. storočia. Ed. Mária Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004, s. 150 – 158.; Mária ŠIMKOVÁ: Počítačové spracovanie prirodzeného jazyka a Slovenský národný korpus. In: Počítačová podpora prekladu. Zborník prednášok. Ed. Marián Smolík – Jaroslav Šoltys – František Tomášik. Bratislava: Slovenská spoločnosť prekladateľov odbornej literatúry 2003, s. 15 – 19.
resourceType	corpus
mediaType	text
lingualityType	monolingual

languageId	sk
size	7 700 000
sizeUnit	tokens

resourceName	Slovak Web Corpus
resourceShortName	sk-web
downloadLocation	
dateCreation	2011
projectPartner	LSIL
iprHolder.organizationName	LSIL/various
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available-restricted use
license	other
resourceLocation	LSIL
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	human use / NLP applications
actualUse	human use / NLP applications
description	Web corpus contains texts downloaded from the .sk domain. The texts are automatically lemmatized and morphologically tagged.
relevantPublications	-
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	sk
size	900,000,000
sizeUnit	token

resourceName	Slovak Legal TextsCorpus
resourceShortName	legal
downloadLocation	
dateCreation	2011
projectPartner	LSIL
iprHolder.organizationName	LSIL; MS SR
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available-restricted use
license	other
resourceLocation	LSIL

distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	human use / NLP applications
actualUse	human use / NLP applications
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	-
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	sk
size	146 000 000
sizeUnit	token

resourceName	Slovak-Czech Parallel Corpus
resourceShortName	sk-cs
downloadLocation	http://korpus.juls.savba.sk/skcs.html
dateCreation	2010
projectPartner	LSIL
iprHolder.organizationName	LSIL/various
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available-restricted use
license	other
resourceLocation	LSIL
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	human use / NLP applications
actualUse	human use / NLP applications
description	Parallel Slovak-Czech corpus is a corpus of sentence aligned texts, mostly fiction. The Slovak texts are morphologically annotated and disambiguated using the system applied in the Slovak National Corpus, Czech texts are annotated with the morče tagger. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
relevantPublications	-
resourceType	corpus
mediaType	text
lingualityType	bilingual
languageId	sk; cs
size	730 000
sizeUnit	sentence

resourceName	Slovak-English Parallel Corpus
resourceShortName	sk-en
downloadLocation	http://korpus.juls.savba.sk/sken.html
dateCreation	2011
projectPartner	LSIL
iprHolder.organizationName	LSIL
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available-restricted use
license	other
resourceLocation	LSIL
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	human use / NLP applications
actualUse	human use / NLP applications
description	The corpus consists of parallel Slovak and English texts (mostly fiction), with automatic lemmatization, morphological analysis (for Slovak), POS tagging (for English). The corpus consists of original English language books and their Slovak translations. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.
relevantPublications	-
resourceType	corpus
mediaType	text
lingualityType	bilingual
languageId	sk; en
size	1 500 000
sizeUnit	sentence

resourceName	Slovak-French Parallel Corpus
resourceShortName	sk-fr
downloadLocation	http://korpus.juls.savba.sk/frask/
dateCreation	2007
projectPartner	LSIL
iprHolder.organizationName	LSIL/various
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available-restricted use
license	other
resourceLocation	LSIL
distributionAccessMedium	accessibleThroughInterface

restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	human use / NLP applications
actualUse	human use / NLP applications
description	The corpus contains original French fiction texts and their Slovak translations, with automatically aligned sentences.
relevantPublications	Dorota VASILIŠINOVÁ, Radovan GARABÍK: 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.
resourceType	corpus
mediaType	text
lingualityType	bilingual
languageId	sk; fra
size	21 000
sizeUnit	sentence

resourceName	Slovak-Russian Parallel Corpus
resourceShortName	sk-ru
downloadLocation	http://korpus.juls.savba.sk/parus/
dateCreation	2006
projectPartner	LSIL
iprHolder.organizationName	LSIL/various
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available-restricted use
license	other
resourceLocation	LSIL
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	The corpus contains original Russian fiction texts and their Slovak translations, with automatically aligned sentences.
relevantPublications	Radovan Garabík: Захаров, Виктор Павлович: Параллельный русско-словацкий корпус. In: Труды международной конференции Корпусная лингвистика. Санкт-Петербург: Издательство С.-Петербургского университета 2006, p. 81 – 87.
resourceType	corpus
mediaType	text
lingualityType	bilingual
languageId	sk; rus

size	100 000
sizeUnit	sentence

resourceName	Corpus of Spoken Slovak
resourceShortName	hovor
downloadLocation	http://korpus.juls.savba.sk/shk.html
dateCreation	2008-12-29
projectPartner	LSIL
iprHolder.organizationName	Slovak National Corpus
contact.Person.surname	Gajdošová
contact.Person.givenName	Katarína
contact.Person.email	katarinag@korpus.juls.savba.sk
DistributionInfo	available-unrestricted use
license	CC-BY-SA, GFDL v1.3, Affero GPL v3
resourceLocation	LSIS
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	attribution, shareAlike
licenseSignatory.Person.position	director
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	The database of the Corpus of Spoken Slovak contains audio records of spontaneous and semi-prepared speech from the entire Slovak territory and their text transcripts. Specific characteristics of spoken language are selectively captured in the transcripts, such as irregular structure of an utterance, pronunciation variants, means of speech modulation, and the presence of non-linguistic elements. The Corpus of Spoken Slovak provides material for research and description of the real form of contemporary standard spoken Slovak.
relevantPublications	Radovan GARABÍK, Milan RUSKO: 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.; Radovan GARABÍK, Agáta KARČOVÁ, Mária ŠIMKOVÁ, Katarína GAJDOŠOVÁ: Hovorený korpus slovenčiny. In: Čeština v mluveném korpusu. Ed. M. Kopřivová – M. Waclawičová. Praha: Nakladatelství Lidové noviny 2008, s. 227 – 233.; Katarína GAJDOŠOVÁ: Využitie a výslovnosť skratiek v Slovenskom hovorenom korpuse. Prednesený na konferencii Slovakistika vo všeobecnolingvistickej perspektíve organizovanej FF UPJŠ v Košiciach v dňoch 28. – 29. 5. 2009. V tlači.; Katarína GAJDOŠOVÁ: Cudzojazyčné výrazy v Slovenskom hovorenom korpuse. In: Slovo o slove 16. Prešov: Pedagogická fakulta Prešovskej univerzity v Prešove 2010, s. 190 – 197.; Katarína GAJDOŠOVÁ: Metadáta v Slovenskom hovorenom korpuse. In: VARIA XVII. Bratislava: Slovenská jazykovedná spoločnosť pri SAV – FF KU v Ružomberku 2010. s. 115 – 120.
resourceType	corpus
mediaType	audio text
lingualityType	monolingual
languageId	sk

size	178 (audio), 1 643 000 (text)
sizeUnit	hours (audio), tokens (text)

resourceName	Slovak Morphology Database (Lexicon)
resourceShortName	ma
downloadLocation	https://data.juls.savba.sk/ma/
dateCreation	2005
projectPartner	LSIL
iprHolder.organizationName	LSIL/various
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available-unrestricted use
license	CC-BY-SA, GFDL v1.3, Affero GPL v3
resourceLocation	LSIL
distributionAccessMedium	downloadable
restrictionsOfUse	attribution, shareAlike
licenseSignatory.Person.position	director
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	Slovak Morphological Database is a database of lemmas and their inflected wordforms with MSD tags
relevantPublications	Radovan GARABÍK: Levenshtein Edit Operations as a Base for a Morphology Analyzer. In: Computer Treatment of Slavic and East European Languages. Proceedings of the conference Slovko 2005. Ed. R. Garabík. Bratislava: Veda 2005.; Radovan GARABÍK, Lucia GIANITSOVÁ, Lucia OLOŠTIAKOVÁ: Manual Morphological Annotation of the Slovak Translation of Orwell's Novel 1984 – Methods and Findings. In: Computer Treatment of Slavic and East European Languages. Proceedings of the conference.; Radovan GARABÍK: Slovak morphology analyzer based on Levenshtein edit operations. In: Proceedings of the WIKT'06 conference, Bratislava 2006, p. 2 – 5.; Radovan GARABÍK: Storing Morphology Information in a Wiki. In: Lexicographic Tools and Techniques. Moscow: IITP RAS 2008, p. 55 – 59.
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	monolingual
languageId	sk
size	77000
sizeUnit	lemma

resourceName	Slovak Terminology Database
resourceShortName	STD
downloadLocation	https://data.juls.savba.sk/std/
dateCreation	2008

projectPartner	LSIL
iprHolder.organizationName	Slovak National Corpus
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available
license	CC-BY-SA, GFDL v1.3, Affero GPL v3
resourceLocation	LSIL
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	attribution, shareAlike
licenseSignatory.Person.position	director
foreseenUse	human use
actualUse	human use
description	
relevantPublications	Jana Levická: Teoretické východiská budovania terminologickej databázy. In: Odborný preklad 2. Ed. J. Šoltýs. Bratislava: AnaPress/Slovenská spoločnosť prekladateľov odbornej literatúry 2006, p. 73 – 81. ISBN 80-89137-54-5
resourceType	
mediaType	
lingualityType	monolingual
languageId	sk
size	4,500
sizeUnit	entries

resourceName	Slovak Treebank
resourceShortName	
downloadLocation	-
dateCreation	2010
projectPartner	LSIL
iprHolder.organizationName	LSIL
contact.Person.surname	Gajdošová
contact.Person.givenName	Katarína
contact.Person.email	katarinag@korpus.juls.savba.sk
DistributionInfo	available-restricted use
license	ANA+NC
resourceLocation	LSIL
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	human use NLP applications
actualUse	human use
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	Mária ŠIMKOVÁ, Radovan GARABÍK: Синтаксическая разметка в Словацком национальном корпусе. In: Труды международной конференции Корпусная лингвистика – 2006. Sankt-Petersburg: St. Petersburg University Press 2006, p. 389 – 394.; Mária ŠIMKOVÁ, Katarína GAJDOŠOVÁ: Slovenský závislostný korpus. In: Gramatika a korpus 2007. Ed. F. Štícha, M. Fried. Praha: Academia 2008. p. 135 – 141.
resourceType	lexicalConceptualResource
mediaType	text
lingualityType	monolingual
languageId	sk
size	50,000
sizeUnit	sentence

resourceName	Slovak WordNet
resourceShortName	wn
downloadLocation	https://data.juls.savba.sk/intranet/wn
dateCreation	2011
projectPartner	LSIL
iprHolder.organizationName	LSIL
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
DistributionInfo	available-unrestricted use
license	CC-BY-SA, GFDL v1.3, Affero GPL v3
resourceLocation	LSIL
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	attribution shareAlike
licenseSignatory.Person.position	director
foreseenUse	human use NLP applications
actualUse	human use NLP applications
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
resourceType	lexicalConceptualResource

mediaType	text
lingualityType	multilingual
languageId	sk; pl; de; lt
size	12,500
sizeUnit	synset