



CESAR

Central and South-East European Resources
Project no. 271022

Deliverable D2.5

Report on resources of further interest

Version No. 1.3
31/07/2012

Document Information

Deliverable number:	D2.5
Deliverable title:	Report on resources of further interest
Due date of deliverable:	31/07/2012
Actual submission date of deliverable:	31/07/2012
Main Author(s):	Svetla Koeva (IBL)
Participants:	Tamás Váradi (HASRIL) Tibor Pintér (HASRIL) Szászák György (BME-TMIT) Radovan Garabík (LSIL) Maciej Ogrodniczuk (IPIPAN) Adam Przepiórkowski (IPIPAN) Piotr Pezik (UŁodz) Marko Tadić (FFZG) Duško Vitas (UBG) Cvetana Krstev (UBG) Ivelina Stoyanova (IBL)
Internal reviewer:	Tamás Váradi (HASRIL)
Workpackage:	2
Workpackage title:	Analysis and selection of language resources
Workpackage leader:	IBL
Dissemination Level:	Public
Version:	1.3
Keywords:	language resources, tools for natural language processing, language technologies

History of Versions

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

EXECUTIVE SUMMARY

The deliverable gives a detailed description on the identified resources that are of further interest to the consortium. The document summarises the language resources (language by language) gathered in this task of the project. A more detailed description of the resources is given in the annex.

Table of Contents

1. Background	4
1.1. Project objectives	4
1.2. Baseline situation	4
1.3. Target resources and users	5
2. A common and shared resource description	6
3. Resources identified via CESAR until month eighteenth	7
4. Resources of further interest identified via CESAR	8
4.1. Summary of the language resources of further interest developed in Bulgaria	8
4.2. Summary of the language resources of further interest developed in Croatia	9
4.3. Summary of the language resources of further interest developed in Hungary	10
4.4. Summary of the language resources of further interest developed in Poland	12
4.5. Summary of the language resources of further interest developed in Serbia	13
4.6. Summary of the language resources of further interest developed in Slovakia	15
5. Resources of further interest – brief summary	17
6. Annex 1 – detailed description of resources of further interest	18
6.1. Bulgarian language resources of further interest - detailed specification	18
6.2. Croatian language resources of further interest - detailed specification	23
6.3. Hungarian language resources of further interest - detailed specification	25
6.4. Polish language resources of further interest - detailed specification	32
6.5. Serbian language resources of further interest - detailed specification	38
6.6. Slovak language resources of further interest – detailed specification	43
7. Annex 2 – information for all identified language resources	48
7.1. Bulgarian	48
7.2. Croatian	51
7.3. Hungarian	55
7.4. Polish	63
7.5. Serbian	69
7.6. Slovak	73

1. Background

1.1. Project objectives

The CESAR project aims to establish positive practices in addressing the needs of human language technologies (crucially depending on language resources and tools). Central activities related to the outlined aim include developing, enhancing, upgrading, standardising and cross-linking a wide variety of language resources and tools. Beside the central activities it is also important to provide extensive access to the language technologies for a wide range of professionals which will contribute to the establishment of an open linguistic infrastructure.

The main goals of the CESAR project are:

- to provide a description of the national (resp. language community) landscape in terms of language use; language-savvy products and services, language technologies and resources; main actors (research, industry, government and society); public policies and programmes; prevailing standards and practices; current level of development;
- to contribute to a pan-European digital resource exchange facility by collecting resources and by documenting, linking and upgrading them to agreed standards and guidelines;
- to help build and operate broad, non-commercial, community-driven, interconnected repositories, exchanges, facilities, etc. that can be used by language researchers, developers and professionals;
- to 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 builds on existing resources and previous national or international activities and on that basis creates a comprehensive language-resource platform enabling and supporting large-scale multi- and cross-lingual products and services. The CESAR project has a strong focus on six Central and South-East European languages, all of them considered less-resourced: Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak. Four of these languages are the official languages of EU member states (Bulgarian, Hungarian, Polish, Slovak) while the other two (Croatian and Serbian) are languages of states which are scheduled to join the EU in the near future. The development of linguistic resources and technologies for these languages will in cooperation with the META-NET ensure their interoperability and compliance with widely acknowledged standards.

In the frame of these tasks, language resources and tools already developed or still under development have been identified. The D2.5 Report on resources of further interest to the consortium represents the resources for Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak, which are of continued interest to the consortium with respect to enlarging and enhancing the development of language technologies in Europe.

1.3. Target resources and users

The CESAR targets a wide range of users including researchers in the field of linguistics, natural language processing, computer science, developers and other professionals involved in industry, private and public institutions, national governments, EU institutions.

A large variety of resources are considered essential for the development of language technologies – large written and spoken corpora, annotated with various linguistic information, both monolingual and multilingual, general and specialised lexical resources, grammars, ontologies, processing and annotation tools and technologies. The available resources are constantly extended and enhanced with new linguistic information, while new resources are continuously being developed.

2. A common and shared resource description

The CESAR supports the goal of a common and shared resource description between the four projects constituting META-NET (CESAR, METANET4U, META-NORD, and T4ME).

The purpose was to gather information about the resources of further interest to the consortium. These include both data (textual or multimodal, monolingual or multilingual corpora, lexical resources, etc.) and technologies (tools and services). The metadata provides a basic description of the resource and its actual or upcoming state, as well as the data types and media type of the resource or the language covered by it.

The unified approach towards resource description will facilitate the application of resources for various research tasks and will enhance their exploitation. The META-SHARE metadata are descriptions of Language Resources, encompassing both data sets (textual, multimodal/multimedia and lexical data, grammars, language models etc.) and tools/technologies/services used for their processing.

The META-SHARE metadata group together semantically coherent elements and relations as well as other components. The elements encode descriptive features of language resources, the relations link together resources that are included in the META-SHARE, as well as resources with related entities (e.g. documentation manuals, publications, licences etc.). The elements show two basic levels of description: an initial level providing the basic elements for the description of a resource (minimal schema), and a higher degree of granularity (maximal schema), providing more detailed information.

The specifications used for the description of the language resources at the D2.3 Deliverable - report on resources (actually and potentially) available to the consortium as well as at the D2.5 Report on resources of further interest, are the common META-SHARE specifications (closely related with the minimal schema). The goal is to unify the description of the identified language resources as well as to provide the most important information for them.

3. Resources identified via CESAR until M18

Complying with some of the main goals of the CESAR project D2.3 provides a detailed description of the national (resp. language community) landscape in terms of language-savvy products and services and language technologies and resources.

To contribute to a pan-European digital resource exchange facility a number of resources were collected, documented, linked and upgraded to agreed standards and guidelines. Thus broad, non-commercial, community-driven, interconnected repositories, exchanges, facilities were built and operated, so that they can be used by language researchers, developers and professionals.

An overview of the outcomes within D2.3 shows that during the period of the project work package (18 months) a total of 213 resources altogether were identified as actually or potentially available to the consortium. Approximately half of the resources are corpora (87 text and 32 audio or multimedia). 51 of the resources are lexical/conceptual databases while 53 are technology tools or services. 47 of the resources (approximately 1/4) are multilingual (covering the different languages).

Finally, 94 of the resources are identified outside the consortium (which makes approximately 44% of the gathered resources).

This brief overview of D2.3 presents enough evidence for the achievement of the project goals, namely to bridge the technological gap between the Central European region and other parts of Europe by filling some obvious and important gaps in language resources and tools infrastructure.

Resources per Country	Total	By Resource type				By Linguality			Outside the consortium
		Text Corpora	Audio Corpora	Lexical / Conceptual Database	Technology tool / service	Mono lingual	Bilingual	Multi lingual	
Bulgaria	29	7	2	6	14	20	2	7	6
Croatia	20	11	1	2	6	14	1	5	3
Hungary	56	19	17	9	11	50	5	1	29
Poland	54	16	7	18	13	42	5	7	33
Serbia	32	14	4	8	6	23	5	4	14
Slovakia	22	10	1	8	3	17	3	2	9
Total	213	87	32	51	53	166	21	26	94

Table 1. Summary of the identified language resources

4. Resources of further interest identified via CESAR

4.1. Summary of the language resources of further interest developed in Bulgaria

The resources developed in Bulgaria which may be of further interest for the community:

- **Diachronic corpus of Bulgarian Language (histdict)** – Corpus of Medieval and Early Modern Bulgarian texts and manuscripts.
- **Bugarian Framenet (BulFrameNet)** – The Bulgarian FrameNet represents general semantic and language-specific lexical-semantic and syntactic combinatory properties of Bulgarian lexical units. BulFrameNet database so far contains unique descriptions of over 3 000 Bulgarian lexical units, approx. one tenth of them aligned with appropriate semantic frames.
- **Bulgarian Grammatical Dictionary (BulGram)** - The Grammar Dictionary of Bulgarian is an inflexion dictionary that consists of approximately 85 000 lemmas and allows automatic word form analysis and generation resulting in approximately 1 mln. 140 thousand word forms. The formal structure of the dictionary is based on finite state transducers (FSTs) that are widely applied in up-to-date electronic dictionaries.
- **Translation Reference Library (TREFL)** – 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.
- **Real Time Comparison (RTComp)** – 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.
- **Bulgarian Word Sense Disambiguation Tool** currently uses 5 independent weak classifiers and an ensemble one that combines all of them. Each of the 6 classifiers provides confidence distribution over the senses for a particular single word or MWE. The current version outperforms the calculated random sense baseline by 24 points.

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	Linguality Type	Outside the consortium
Diachronic corpus of Bulgarian Language	histdict	http://histdict.uni-sofia.bg	corpus			multilingual	yes
Bulgarian FrameNet	BulFrameNet	http://dcl.bas.bg/LexIt	lexicalConceptualResource	3 000	concets	multilingual	no
Bulgarian Grammatical Dictionary	BulGram	http://dcl.bas.bg/est	lexicalConceptualResource	85 000	lemma	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
RTComp - Real Time Comparison	RTComp	http://web.uni-plovdiv.bg/rousni/rtcomp	technology tool	10 MB	file	multilingual	yes
Bulgarian Word Sense Disambiguation Tool	BulWSD	http://dcl.bas.bg/en/programs_en.html	technologyToolService	-	-	monolingual	no

Table 2. Summary of the language resources of further interest developed in Bulgaria

4.2. Summary of the language resources of further interest developed in Croatia

The resources developed in Croatia which may be of further interest for the community:

- **Croatian Speech Corpus (CroSpeak Corpus)** – Croatian Speech Corpus (CroSpeak Corpus) is the corpus of recorded Croatian speech covering radio weather forecasts, radio news, read tales, weather dialogs, and TV news (featuring unspontaneous and spontaneous speech). Overall size of the corpus is 19.35 hours and 227,280 tokens. All utterances have been transcribed following standard Croatian orthography. The corpus has been compiled and processed at Department of Information Sciences, University of Rijeka.
- **Croatian Academic Spelling Checker** – Croatian Academic Spelling Checker (Hascheck) is one of the oldest Internet services in Croatia. In various forms it acts as a public service and free spelling checker for text written in Croatian language since spring 1994. Hascheck's dictionary database is organized into three sections: (1) Croatian general lexicon, (2) Croatian

lexicon of names, (3) English general lexicon. Dictionary database is not static and it is being constantly improved.

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	Linguality Type	Outside the consortium
Croatian Speech Corpus	CroSpeak Corpus	http://www.inf.uniri.hr/~ivoi/CROSP/EECH/index.htm	corpus	227 280	token	monolingual	yes
Croatian Academic Spelling Checker	Hascheck	http://hachek.tel.fer.hr/	technologyToolService			monolingual	yes

Table 3. Summary of the language resources of further interest developed in Croatia.

4.3. Summary of the language resources of further interest developed in Hungary

The resources developed in Hungary which may be of further interest for the community:

- **Tesztel Hungarian Noisy Telephone Speech Corpus** - Noisy speech samples for noise robust ASR in Hungarian.
- **A Hungarian Child Database for Speech Processing Applications** - Child speech for Hungarian, useful for speech training and language learning applications for children.
- **ht-online** is a unique lexical database of the most common loanwords in Hungarian language used outside Hungary (collected from 7 regions). The database should be used as special lexical resource in the Hungarian language tools based on the Hungarian morphology.
- **Hungarian Concise Dictionary (with sample sentences)** is a unique dictionary of Hungarian language covering 16,000 headwords (entries) followed by frequency data. Each entry describes the most common forms (selected on pragmatical basis) of the headword. The entries are divided into meanings – up to 33,000 carefully selected and stylistically labeled meanings. The dictionary contains sentences brought from real language use and 3,000 phrasems.
- **Child Language Corpus** consists of 60 interviews with 4/6-5/6 year-old Hungarian children (from Budapest and from different socio-economical backgrounds) with more than 30 hours recording. The interviews include several tasks (picture-based story-telling, telling the rules of a well-known game) and guided conversations. Each interview was conducted by an adult tester. The resource is available in chat (CHILDES) transcription format.

- **BABEL and MRBA sentence modality annotation for Hungarian** - Corpus holding modality annotations for subparts of Hungarian BABEL and MRBA corpora. If extended, a useful source for research and analysis or recognition of sentence modality.
- **Multilingual speech segmentation tool** - Multilingual speech segmentation tool.
- **Sentence modality recognizer** - Sentence modality recognizer from speech for Hungarian and German, can be used in speech recognition and understanding.

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	Linguality Type	Outside the consortium
Tesztel Hungarian Noisy Telephone Speech Corpus		BME-TMIT	corpus	100	speakers	monolingual	no
A Hungarian Child Database for Speech Processing Applications		BME-TMIT	corpus	72	speakers	monolingual	no
ht-online	ht-online	Termini Research Network	lexical / conceptual resource	4000	entry	monolingual	yes
Hungarian concise dictionary (with sample sentences)	HCD	TINTA Publishing House	lexical / conceptual resource	16000	entry	monolingual	yes
Child language corpus	CHILC	HASRIL	corpus	60	interview	monolingual	no
BABEL and MRBA sentence modality annotation for Hungarian		BME-TMIT	corpus	50	speakers	monolingual	no
Multilingual speech segmentation tool		BME-TMIT	tool			multilingual	no

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	Linguality Type	Outside the consortium
Sentence modality recognizer		BME-TMIT	tool			bilingual	no

Table 4. Summary of the language resources of further interest developed in Hungary.

4.4. Summary of the language resources of further interest developed in Poland

The resources developed in Poland which may be of further interest for the community:

- o **Polish Radio Żak and Radio Łódź Speech Corpus (RadioZakŁódź)** - A monolingual corpus which at present contains 50 000 words of text and audio.
- o **Świgr** - Świgr is a Prolog parser implementing Świdziński's Formal Grammar of Polish. Świgr uses a bottom-up parsing strategy, which for Polish proved to be superior to the top-down strategy. The parser builds a shared parse forest, which is not only the result but also a means of avoiding unnecessary recomputation. The rules of the grammar are not interpreted at the runtime but they are compiled to Prolog clauses.
- o **Formal Grammar of Polish (GFJP)** is the most extensive and most detailed formal grammar of Polish expressed as a metamorphosis grammar with several extensions, e.g., allowing for permuting phrases. Syntactic units are represented by terms of parameters formalizing various grammatical features of those units. Rules of the grammar define particular units as sequences of other units and establish correspondences between grammatical features (unification). Agreements are accounted for by parameter matching used an extensive set of parameters. The values a given unit is assigned, be it from the top ("syntactic" features) or from the bottom ("lexical" features), to spread down the syntactic tree, reaching most of its constituents. Rules defining different syntactic units (sentences or phrases) follow one format. The grammar has the ambition to define the whole language to cover most structures of Polish.
- o **Anotatornia** - Anotatornia is a tool for the manual on-line annotation of corpora at various linguistic levels. The levels currently implemented are: word-level and sentence-level segmentation, morphosyntax, word sense disambiguation.
- o **Ruler** - Ruler is a rule-based co-reference resolver for Polish. The implemented module uses standard best-first entity-based model based on syntactic constraints (elimination of nested nominal groups), syntactic filters (elimination of syntactic incompatible heads), semantic filters (wordnet-derived compatibility) and selection (weighted scoring).

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

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	Linguality Type	Outside the consortium
Polish Radio Zak and Radio Łódź Speech Corpus	RadioZakŁódź	http://www.zak.lodz.pl/ , http://www.radiolodz.pl/	corpus	50 000	word	monolingual	yes
Świgr	Świgr	–	toolService	–	–	monolingual	yes/no
Formal Grammar of Polish	GFJP	–	lexicalConceptualResource	460	rules	monolingual	yes/no
Anotatornia	Anotatornia	–	toolService	–	–	monolingual	yes/no
Ruler	Ruler	–	toolService	–	–	monolingual	yes/no
PolSumm	PolSumm	–	toolService	–	–	monolingual	yes

Table 5. Summary of the language resources of further interest developed in Poland.

4.5. Summary of the language resources of further interest developed in Serbia

The resources developed in Serbia which may be of further interest for the community:

- o **Terminological Database for Geology (GeolISSTerm)** - The electronic dictionary of geologic terms is a special-purpose taxonomy of basic geologic concepts and terms. GeolISSTerm is an elementary electronic resource in the process of domain formation in the Geologic Information System of Serbia (GeolISS). It is the core of GeolISS through which validation, classification and specification of attributes of the observed and the interpreted takes place.
- o **Emotion classification of Serbian Texts** - This system for emotion classification of Serbian texts is based on an ontology built specially for this purpose that functions as an emotion classifier. The application is realized on

Csharp Net Framework platform. It can be tested on texts in .html and .txt formats and it accepts both Cyrillic and Latin scripts. Text files can be manually pasted, uploaded from a local system or used directly from a given URL address on Web.

- **Named entities module for Serbian (SrpNE-module)** - This module for named entity recognition and tagging is based on Serbian morphological e-dictionaries and a large collection of Finite-State Transducers (in the form of cascades). It recognizes and tags: persons, person roles and functions, temporal expressions, mount expressions (including measures and money expressions) and organizations. The module is integrated in a web service and tags NEs in texts uploaded by users.
- **A web tool for aligned text search** - This is a web tool for effective search of aligned and annotated texts. It is especially designed for texts in which named entities were tagged. Its purpose is to compare annotation of NEs in aligned text and for that purpose a language independent classification schema for NEs is used.
- **Web applications (NE extraction from web pages)** - This is a web tool for extraction of proper names from categories given in Wikipedia for English, French, Serbian, Polish.
- **Language Model for Serbian** – This language model of Serbian is produced on the basis of the large newspaper corpus (approx. 4 million articles) using the standard methodology for such models.

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	Linguality Type	Outside the consortium
Terminological Database for geology	GeolISSTerm	http://www.rgf.bg.ac.rs/	lexical / conceptual resource	3500	concepts	bilingual	yes
Emotion classification of Serbian Texts		http://korpus.matf.bg.ac.rs	tool			monolingual	no
Named entities module for Serbian	SrpNE-module	http://korpus.matf.bg.ac.rs	tool			monolingual	no
A web tool for aligned text search		http://korpus.matf.bg.ac.rs	tool			multilingual	no
Web applications (NE extraction from web pages)		http://korpus.matf.bg.ac.rs	tool			multilingual	no

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	Linguality Type	Outside the consortium
Language model for Serbian		http://www.arhiv.rs/	lexical resource			monolingual	yes

Table 6. Summary of the language resources of further interest developed in Serbia.

4.6. Summary of the language resources of further interest developed in Slovakia

The resources developed in Slovakia which may be of further interest for the community:

- **Slovak–Russian Parallel Corpus (sk-ru)** – original Russian fiction texts and their Slovak translations, with automatically aligned sentences.
- **Slovak–French Parallel Corpus (sk-fr)** – original French fiction texts and their Slovak translations, with automatically aligned sentences.
- **Database of root morphemes** - The database provides alternative approach to morphology analysis. It contains 67,000 linguistic units with deep morphematic linguistic analysis. It has been compiled at the Prešov University in Prešov and has been used as a basis for a published Slovník koreňových morfém slovenčiny (M. Sokolová et al.). ISBN 9788080683191.
- **Dictionary of Slovak Adjective Collocations** - The dictionary provides an overview of the combinatorial behaviour of words – contains collocation profiles of the most frequent Slovak adjectives. The combinatorial potentials of word forms of a word are the basis for the creation of so-called collocational templates which the patterns of collocations are based on. The dictionary is currently being compiled – at the time of writing it contains collocation profiles of 140 adjectives. The dictionary is being created at the University of St. Cyril and Methodius in Trnava, with input from the L. Štúr Institute of Linguistics.
- **Dictionary of German-Slovak Collocations** - The dictionary provides confrontational overview of the combinatorial behaviour of words in bilingual comparison. The database consists of German collocations (currently 440 profiles) with Slovak equivalents. The dictionary is being created at the University of St. Cyril and Methodius in Trnava.
- **Multimodal multilingual dictionary of gestures (DiGest)** - The dictionary contains database of extra-verbal expressions. Its current version contains several hundreds of gestures represented by a still image, a description of the gesture and its meaning, and optional sound and video records. The current version includes language and culture dependent content for American English, Slovak, Italian, and Mongolian. Entries for Japanese, Chinese, and Hungarian are also included. The database has been compiled at the Institute of Informatics, Slovak Academy of Sciences.

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	Linguality Type	Outside the consortium
Slovak-Russian Parallel Corpus	SK-RU	LSIL	corpus	100000	sentence	bilingual	yes (partly)
Slovak-French Parallel Corpus	SK-FR	LSIL	corpus	21000	sentence	bilingual	no
Database of root morphemes	Database of root morphemes	Prešov University	Lexical / Conceptual Resource	67000	root morpheme	monolingual	no
Dictionary of Slovak Adjective Collocations	Dictionary of Slovak Adjective Collocations	University of St. Cyril and Methodius Trnava	Lexical / Conceptual Resource	70	entry	monolingual	no
Dictionary of German-Slovak	Dictionary of German-Slovak	University of St. Cyril and Methodius	Lexical / Conceptual Resource	40	entry	bilingual	yes
Multimodal multilingual	DiGest	Institute of Informatics, Slovak Academy	Lexical / Conceptual Resource	324	entry	multilingual	yes

Table 7. Summary of the language resources of further interest developed in Slovakia.

5. Resources of further interest – brief summary

During the first and second CESAR batch an impressive number of resources were published under the META-SHARE:

Partner(s)	Number of resources
Bulgaria	16
Croatia	12
Hungary (two partners)	38
Poland (two partners)	37
Serbia	12
Slovakia	13
Total	128

Table 8. Number of resources published under META-SHARE (first and second batch).

More resources are expected to be published at the third CESAR batch. Currently, only 34 from all 213 identified resources are in the list of resources of further interest. The main reasons can be summarised as follows:

- resources are still under development;
- resources are developing outside the consortium;
- licence agreements are still under negotiation.

All of the identified resources are considered for further publishing under the META-SHARE.

6. Annex 1 – detailed description of resources of further interest

6.1. Bulgarian language resources of further interest - detailed specification

resourceName	<i>Diachronic corpus of Bulgarian Language</i>
resourceShortName	histdict
iprholder.organizationShortName	
contactPerson.surname	Totomanova
contactPerson.givenName	Anna-Maria
contactPerson.email	atotomanova@abv.bg
DistributionInfo	avaialable-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 of 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

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

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

resourceTitle	<i>Bulgarian Grammatical Dictionary</i>
resourceName	BulGram
urlDownload	http://dcl.bas.bg/est/dict.php
dateCreation	1999-2003

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

resourceName	<i>TREFL – Translation Reference Library</i>
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, <i>Théorie, Littérature Epistémologie</i> , 25, pp 105-123 1. Roussi Nikolov & Malina DITCHEVA, Една програма-помощник за превод, четене и писане, Plovdiv University "Paissii Hilendarski" - Bulgaria, <i>Scientific Works</i> , 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	<i>RTComp - Real Time Comparison</i>
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	<i>Bulgarian word sense disambiguation tool</i>
resourceShortName	BulWSD
downloadLocation	
dateCreation	2012
projectPartner	Institute for Bulgarian language
iprHolder.organizationName	Institute for Bulgarian language
contact.Person.surname	Koeva
contact.Person.givenName	Svetla
contact.Person.email	svetla@dcl.bas.bg
DistributionInfo	underNegotiation
license	
resourceLocation	http://dcl.bas.bg/en/programs_en.html
distributionAccessMedium	downloadable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse	nlpApplications
actualUse	nlpApplications

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

6.2. Croatian language resources of further interest - detailed specification

resourceName	<i>Croatian Speech Corpus</i>
resourceShortName	CroSpeak Corpus
downloadLocation	http://www.inf.uniri.hr/~ivoi/CROSPEECH/index.htm
dateCreation	2011
projectPartner	FFZG
iprHolder.organizationName	University of Rijeka
contact.Person.surname	Ipšić
contact.Person.givenName	Ivo
contact.Person.email	ivoi@inf.uniri.hr
DistributionInfo	available-restricted
license	
resourceLocation	http://www.inf.uniri.hr/~ivoi/CROSPEECH/index.htm
distributionAccessMedium	not yet available for internet access
restrictionsOfUse	following the license

licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications
description	Croatian Speech Corpus (CroSpeak Corpus) is the corpus of recorded Croatian speech covering radio weather forecasts, radio news, read tales, weather dialogs, and TV news (featuring unspontaneous and spontaneous speech). Overall size of the corpus is 19.35 hours and 227,280 tokens. All utterances have been transcribed following standard Croatian orthography. The corpus has been compiled and processed at Department of Information Sciences, University of Rijeka
relevantPublications	Martinčić-Ipšić, S., Pobar, M., Ipšić, I. Croatian Large Vocabulary Automatic Speech Recognition, <i>Automatika</i> , Vol 52, no 2 (2011), p. 147-157. Martinčić-Ipšić, S., Ipšić, I. Recognition of Croatian Broadcast Speech. Budin, L. (ed.), Ribarić, S., (ed.). XXVII. MIPRO 2004, Opatija, Vol. CTS + CIS, p. 111-114. 2004.
resourceType	corpus
mediaType	speech
lingualityType	monolingual
languageId	hrv
size	227 280
sizeUnit	token

resourceName	<i>Croatian Academic Spelling Checker</i>
resourceShortName	Hascheck
downloadLocation	http://hacheck.tel.fer.hr/
dateCreation	1994
projectPartner	FFZG
iprHolder.organizationName	FER
contact.Person.surname	Dembitz
contact.Person.givenName	Šandor
contact.Person.email	sandor.dembitz@fer.hr
DistributionInfo	available-unrestricted use
license	
resourceLocation	http://hacheck.tel.fer.hr/
distributionAccessMedium	web service
restrictionsOfUse	following the license restrictions
licenseSignatory.Person.position	
foreseenUse	human use; NLP applications
actualUse	human use; NLP applications

description	Croatian Academic Spelling Checker (Hascheck) is one of the oldest Internet services in Croatia. In various forms it acts as a public service and free spelling checker for text written in Croatian language since spring 1994. Hascheck's dictionary database is organized into three sections: (1) Croatian general lexicon, (2) Croatian lexicon of names, (3) English general lexicon. Dictionary database is not static and it is being constantly improved. In the background a working expert system that learns new words from texts submitted for processing. The database is maintained through supervised learning process and currently it exceeds one million types, which all have been attested in the texts written in Croatian language. The core solutions for spelling checker were written by Šandor Dembitz and the majority of web site processing was written by Gordon Gledec, while the web interface was written by Hrvoje Miholić.
relevantPublications	Dembitz, Š., Knežević, P., Sokele, M. Developing a Spell Checker as an Expert System. // CIT. Journal of computing and information technology. 11 (2003) , 4; 285-292.
resourceType	technologyToolService
mediaType	text
lingualityType	multilingual
languageId	hrv, eng
size	-
sizeUnit	-

6.3. Hungarian language resources of further interest - detailed specification

resourceName	<i>Tesztel Hungarian Noisy Telephone Speech Corpus</i>
resourceShortName	Not applicable
downloadLocation	if applicable
dateCreation	2006-12-31
projectPartner	BME-TMIT
iprHolder.organizationName	BME-TMIT
contact.Person.surname	Szaszák
contact.Person.givenName	György
contact.Person.email	szaszak@tmit.bme.hu
DistributionInfo	notAvailable
license	notAvailable
resourceLocation	BME-TMIT
distributionAccessMedium	DVD-R

restrictionsOfUse	informResourceOwner academic-nonCommercialUse attribution shareAlike
licenseSignatory.Person.position	Head of dept.
foreseenUse	human use nlpApplications
actualUse	human use nlpApplications
description	<p>This database contains noisy speech samples for noise robust ASR in Hungarian, recorded via PSTN and mobile telephone network. The aim of this project was to create a mobile phone voice based Hungarian speech database recorded in noisy environments for testing purposes (also called Tesztel). The database contains voices of 100 speakers, recorded through mobile telephone in noisy environments. The main goal was to test phoneme based recognizers, which have been already trained, so the corpus must have been compact and had to cover as good as possible the specific character of the Hungarian language. The text that the speaker had to tell was designed to contain at least one of every Hungarian phoneme, taking in consideration the statistics of phonemes, diphones, triphones and syllables in Hungarian language.</p> <p>The corpus contains not only continuously told sentences, but command words, spelled forenames, numbers, dates, different currency types, city names, questions with yes/no answer, phonetically rich words. The database contains mostly spontaneous speech. Since the whole database contains speech recorded in noisy environments, we wanted to find out an average value of the signal to noise ratio for the recorded speech.</p>
relevantPublications	
resourceType	corpus
MediaType	audio
lingualityType	monolingual
languageId	hu
size	100
sizeUnit	speaker

resourceName	<i>Hungarian Child Database for Speech Processing Applications</i>
resourceShortName	Not applicable
downloadLocation	if applicable
dateCreation	
projectPartner	BME-TMIT
iprHolder.organizationName	BME-TMIT
contact.Person.surname	Szaszák
contact.Person.givenName	György
contact.Person.email	szaszak@tmit.bme.hu

DistributionInfo	notAvailable
license	notAvailable
resourceLocation	BME-TMIT
distributionAccessMedium	DVD-R
restrictionsOfUse	informResourceOwner academic-nonCommercialUse attribution shareAlike
licenseSignatory.Person.position	Head of dept.
foreseenUse	human use nlpApplications
actualUse	human use nlpApplications
description	Child speech for Hungarian, useful for speech training and language learning applications for children. More detailed description of the corpus was uploaded to: http://alpha.tmit.bme.hu/speech/paperc013.php
relevantPublications	Vicsi, K. - Csatári, F. - Bakcsi, Zs. "Distance Score evaluation of the visualized speech spectra at audio-visual articulation training" - EUROSPEECH. Vicsi, K. - Roach, P. - Öster, A. - Kacic, Z. - Barczikay, P. - Sinka, I. : SPECO, a multimedia multilingual teaching and training system for speech handicapped children EUROSPEECH '99
resourceType	corpus
MediaType	audio
lingualityType	monolingual
languageId	hu
size	72
sizeUnit	speaker

resourceName	<i>ht-online</i>
resourceShortName	ht-online
downloadLocation	ht.nytudhu/htonline
dateCreation	2009-2012
projectPartner	HASRIL
iprHolder.organizationName	Termini Research Centre
contact.Person.surname	Pintér
contact.Person.givenName	Tibor
contact.Person.email	pinter.tibor@nytud.mta.hu
DistributionInfo	available-unrestricted use
license	XXX
resourceLocation	HASRIL
distributionAccessMedium	accessibleThroughInterface

restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	János Péntek
foreseenUse	NLP applications
actualUse	human use
description	A unique lexical database of the most common loanwords in Hungarian language used outside Hungary (collected from 7 regions). The database should be used as special lexical resource in the Hungarian language tools based on the Hungarian morphology.
relevantPublications	-
resourceType	lexical / conceptual resource
mediaType	text
lingualityType	monolingual
languageId	HU
size	4000
sizeUnit	entries

resourceName	<i>Hungarian concise dictionary (with sample sentences)</i>
resourceShortName	HCD
downloadLocation	–
dateCreation	2011
projectPartner	HASRIL
iprHolder.organizationName	TINTA Publishing House
contact.Person.surname	Pintér
contact.Person.givenName	Tibor
contact.Person.email	pinter.tibor@nytud.mta.hu
DistributionInfo	underNegotiation
license	--
resourceLocation	HASRIL
distributionAccessMedium	hardDisk
restrictionsOfUse	other
licenseSignatory.Person.position	Gábor Kiss
foreseenUse	please, leave the appropriate human use NLP applications
actualUse	please, leave the appropriate human use NLP applications

description	A unique dictionary of Hungarian language of 16 000 headwords (entries) followed by frequency data. Each entry describes the most common forms (given by pragmatical reasons) of the headword. The entries are divided into meanings which counts 33 000 carefully selected and stylistically labelled meanings. The dictionary contains sentences brought from real language use and 3000 phrasems.
relevantPublications	--
resourceType	lexical / conceptual resource
mediaType	text
lingualityType	monolingual
languageId	HU
size	16000
sizeUnit	entries

resourceName	<i>Child language corpus</i>
resourceShortName	CHILC
downloadLocation	--
dateCreation	2012
projectPartner	HASRIL
iprHolder.organizationName	HASRIL
contact.Person.surname	Pintér
contact.Person.givenName	Tibor
contact.Person.email	pinter.tibor@nytud.mta.hu
DistributionInfo	avaialbe-restricted use
license	XXX
resourceLocation	HASRIL
distributionAccessMedium	other
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	Kinga Mátyus
foreseenUse	human use
actualUse	human use

description	The child language corpus consists of 60 interviews with 4;6-5;6 year-old Hungarian children from Budapest, from different socio-economical backgrounds, with more than 30 hours recording. The interviews include several tasks (picture-based story-telling, telling the rules of a well-known game) and guided conversation. Each interview was conducted by an adult experimenter. The resource is available in chat (CHILDES) transcription format.
relevantPublications	-
resourceType	corpus
mediaType	text
lingualityType	monolingual
languageId	HU
size	60
sizeUnit	other

resourceName	<i>BABEL and MRBA sentence modality annotation for Hungarian</i>
resourceShortName	Not applicable
downloadLocation	if applicable
dateCreation	2008-12-31
projectPartner	BME-TMIT
iprHolder.organizationName	BME-TMIT
contact.Person.surname	Szaszák
contact.Person.givenName	György
contact.Person.email	szaszak@tmit.bme.hu
DistributionInfo	notAvailable
license	notAvailable
resourceLocation	BME-TMIT
distributionAccessMedium	DVD-R
restrictionsOfUse	informResourceOwner academic-nonCommercialUse attribution shareAlike
licenseSignatory.Person.position	Head of dept.
foreseenUse	human use nlpApplications
actualUse	human use NlpApplications
description	Corpus holding modality annotations for subparts of Hungarian BABEL and MRBA corpora. If extended, a useful source for research and analysis or recognition of sentence modality.

relevantPublications	Vicsi K, Szaszák Gy: Using prosody to improve automatic speech recognition. SPEECH COMMUNICATION 52:(5) pp. 413-426. (2010)
resourceType	corpus
MediaType	audio
lingualityType	monolingual
languageId	hu
size	50
sizeUnit	speaker

resourceName	<i>Multilingual speech segmentation tool</i>
resourceShortName	Not applicable
downloadLocation	if applicable
dateCreation	2012-12-31
projectPartner	BME-TMIT
iprHolder.organizationName	BME-TMIT
contact.Person.surname	Szaszák
contact.Person.givenName	György
contact.Person.email	szaszak@tmit.bme.hu
DistributionInfo	notAvailable
license	notAvailable
resourceLocation	BME-TMIT
distributionAccessMedium	DVD-R
restrictionsOfUse	informResourceOwner academic-nonCommercialUse attribution noDerivatives
licenseSignatory.Person.position	Head of dept.
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	A multilingual speech segmentation tool, capable of phoneme segmentation of utterances for 6 languages: English, French, Italian, Spanish and Hungarian.
relevantPublications	
resourceType	technology tool / service
MediaType	audio
lingualityType	monolingual
languageId	HU
size	

sizeUnit	other
----------	-------

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

5.4. Polish language resources of further interest - detailed specification

resourceName	<i>Polish Radio Żak and Radio Łódź Speech Corpus</i>
--------------	--

resourceShortName	RadioZakŁódź
downloadLocation	
dateCreation	
projectPartner	Ułódź
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
lingualityType	monolingual
languageId	
size	50 000
sizeUnit	words

resourceName	<i>Świga</i>
resourceShortName	Świga
downloadLocation	–
dateCreation	2003-2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Woliński
contact.Person.givenName	Marcin
contact.Person.email	marcin.wolinski@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	to be defined
resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	NLP applications
actualUse	NLP applications

description	Świgr is a Prolog parser implementing Świdziński's Formal Grammar of Polish. Świgr uses a bottom-up parsing strategy, which for Polish proved to be superior to the top-down strategy. The parser builds a shared parse forest, which is not only the result but also a means of avoiding unnecessary recomputation. The rules of the grammar are not interpreted at the runtime but they are compiled to Prolog clauses.
relevantPublications	not yet available
resourceType	tool
mediaType	text
lingualityType	monolingual
languageId	POL
size	yet unknown
sizeUnit	–

resourceName	<i>Formal Grammar of Polish</i>
resourceShortName	GFJP
downloadLocation	–
dateCreation	1992-2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Świdziński
contact.Person.givenName	Marek
contact.Person.email	m.r.swidzinski@uw.edu.pl
DistributionInfo	available, unrestricted use
license	to be defined
resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	NLP applications
actualUse	NLP applications

description	Formal Grammar of Polish (GFJP) is the most extensive and most detailed formal grammar of Polish expressed as a metamorphosis grammar with several extensions, e.g. allowing for permuting phrases. Syntactic units are represented by terms with parameters formalizing various grammatical features of those units. Rules of the grammar define particular units as sequences of other units and establish correspondences between grammatical features (unification). Agreements are accounted for by parameter matching using an extensive set of parameters. The values a given unit is assigned, be it from the top ("syntactic" features) or from the bottom ("lexical" features), use to spread down the syntactic tree, reaching most of its constituents. Rules defining different syntactic units (sentences or phrases) follow one format. FGP has an ambition to define the whole language – i.e., most structures of Polish are covered.
relevantPublications	not yet available
resourceType	lexical/conceptual resource
mediaType	Text
lingualityType	Monolingual
languageId	POL
size	460
sizeUnit	Rules

resourceName	<i>Anotatornia</i>
resourceShortName	Anotatornia
downloadLocation	http://zil.ipipan.waw.pl/Anotatornia?action=AttachFile&do=view&target=anotatornia-2012-04-10-1206.tgz
dateCreation	2008-2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Lenart
contact.Person.givenName	Michał
contact.Person.email	michal.lenart@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	GPL v. 3
resourceLocation	http://zil.ipipan.waw.pl/Anotatornia/
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	NLP applications
actualUse	NLP applications

description	Anotatornia is a tool for the manual on-line annotation of corpora at various linguistic levels. The levels currently implemented are: word-level and sentence-level segmentation, morphosyntax, word sense disambiguation. Anotatornia implements sophisticated mechanisms of the management of texts, annotators and conflicts.
relevantPublications	Przepiórkowski A., Murzynowski G. Manual annotation of the National Corpus of Polish with Anotatornia. In: Stanisław Goźdz-Roszkowski, ed., Explorations across Languages and Corpora: PALC 2009, pp. 95-103, Frankfurt am Main, 2011. Peter Lang. Hajnicz E., Murzynowski G., Woliński M. ANOTATORNIA – lingwistyczna baza danych. In: Proceedings of the 5th conference InfoBazy 2008, Systems – Applications – Services, pp. 168–173, Sopot, 2008. Centrum Informatyczne TASK, Politechnika Gdańska.
resourceType	tool
mediaType	text
lingualityType	monolingual
languageId	POL
size	–
sizeUnit	–

resourceName	<i>Ruler</i>
resourceShortName	Ruler
downloadLocation	–
dateCreation	2012
projectPartner	IPIPAN
iprHolder.organizationName	IPIPAN
contact.Person.surname	Ogrodniczuk
contact.Person.givenName	Maciej
contact.Person.email	maciej.ogrodniczuk@ipipan.waw.pl
DistributionInfo	available, unrestricted use
license	to be defined
resourceLocation	–
distributionAccessMedium	downloadable (planned)
restrictionsOfUse	attribution, shareAlike (planned)
foreseenUse	NLP applications
actualUse	NLP applications

description	Ruler is a rule-based coreference resolver for Polish. The implemented module uses standard best-first entity-based model based on syntactic constraints (elimination of nested nominal groups), syntactic filters (elimination of syntactic incompatible heads), semantic filters (wordnet-derived compatibility) and selection (weighted scoring). Syntactic properties are obtained from Spejd and its morphological component Morfeusz SGJP which produce NP chunks with detailed morphosyntactic information. Semantic properties are currently based on plWordNet.
relevantPublications	Ogrodniczuk M., Kopeć M. Rule-based coreference resolution module for Polish. In Proceedings of the 8 th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), pp. 191–200. Faro, Portugal.
resourceType	tool
mediaType	text
lingualityType	monolingual
languageId	POL
size	–
sizeUnit	–

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

6.5. Serbian language resources of further interest - detailed specification

resourceName	<i>Terminological Database for Geology</i>
resourceShortName	GeolISSTerm
downloadLocation	-
dateCreation	2006-
projectPartner	University of Belgrade, Faculty of Geology and Mining
iprHolder.organizationName	Ministry of Education and Sciences
contact.Person.surname	Stanković
contact.Person.givenName	Ranka
contact.Person.email	ranka@grf.bg.ac.rs
DistributionInfo	underNegotiation
license	-
resourceLocation	http://www.rgf.bg.ac.rs/
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	-
licenseSignatory.Person.position	-
foreseenUse	NLP applications
actualUse	human use
description	The electronic dictionary of geologic terms (GeolISSTerm) is a special-purpose taxonomy of basic geologic concepts and terms. GeolISSTerm is an elementary electronic resource in the process of domain formation in the Geologic Information System of Serbia (GeolISS). It is the core of GeolISS through which validation, classification and specification of attributes of the observed and the interpreted takes place.
relevantPublications	Stanković, Ranka, and Branislav Trivić, and Olivera Kitanović, and Branislav Blagojević, and Velizar Nikolić. "The Development of the GeolISSTerm Terminological Dictionary." INFOtheca 12, 1: (2011) 49a-63a.
resourceType	lexical / conceptual resource
mediaType	Text
lingualityType	Bilingual
languageId	EN/SR
size	3500

sizeUnit	concepts
----------	----------

resourceName	<i>Emotion classification of Serbian Texts</i>
resourceShortName	
downloadLocation	-
dateCreation	2011-
projectPartner	University of Belgrade, Faculty of Mathematics
iprHolder.organizationName	University of Belgrade, Faculty of Mathematics
contact.Person.surname	Mladenović
contact.Person.givenName	Miljana
contact.Person.email	ml.miljana@gmail.com
DistributionInfo	underNegotiation
license	-
resourceLocation	http://cvetana.miljana.com
distributionAccessMedium	webExecutable
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	-
foreseenUse	NLP applications
actualUse	NLP applications
description	<p>This system for emotion classification of Serbian texts is based on an ontology built specially for this purpose that functions as an emotion classifier. It is based on well-known discrete emotions theories of Arnold, Ekman, Frijda, Gray, Izard, Tomkins, Weiner&Graham, Watson and Plutchik. Each of these theories reviews human emotions as discrete and independent and describes them by small bag of words. These words are used to build the emotions ontology. In order to expand the extraction of information from texts a Serbian associative-dictionary was used coupled with Serbian morphological electronic dictionaries yielding some nine thousand forms used by the system. Extracted RDF structures are then submitted for reasoning and frequencies of emotions are calculated according to each of theories individually. Finally, for the visual presentation of results a separate graphical unit was created. The application is realized on Csharp Net Framework platform. It can be tested on texts in .html and .txt formats and it accepts both Cyrillic and Latin scripts. Text files can be manually pasted, uploaded from a local system or used directly from a given URL address on Web.</p>
relevantPublications	-

resourceType	technology tool / service
mediaType	Text
lingualityType	Monolingual
languageId	SR
size	-
sizeUnit	-

resourceName	<i>Named entities module for Serbian</i>
resourceShortName	SrpNE-module
downloadLocation	-
dateCreation	2009-
projectPartner	University of Belgrade, Faculty of Mathematics
iprHolder.organizationName	University of Belgrade, Faculty of Mathematics
contact.Person.surname	Krstev
contact.Person.givenName	Cvetana
contact.Person.email	cvetana@matf.bg.ac.rs
DistributionInfo	underNegotiation
license	-
resourceLocation	-
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	-
foreseenUse	human use NLP applications
actualUse	human use NLP applications
description	This module for named entity recognition and tagging is based on Serbian morphological e-dictionaries and a large collection of Finite-State Transducers (in the form of cascades). It recognizes and tags: persons, person roles and functions, temporal expressions, mount expressions (including measures and money expressions) and organizations. The module is integrated in a web service and tags NEs in texts uploaded by users.
relevantPublications	Cvetana Krstev, Duško Vitas, Ivan Obradović, Miloš Utvić, "E-Dictionaries and Finite-State Automata for the Recognition of Named Entities", in Proceedings of the 9 th International Workshop on Finite State Methods and Natural Language Processing, FSMNLP 2011, Blois, France, July 12-15, 2010. eds. Andreas Maletti and Matthieu Constant, Association for Computational Linguistics, ISBN 978-3-642-14769-2, pp. 48-56, 2011.
resourceType	technology tool / service
mediaType	text
lingualityType	monolingual
languageId	SR
size	-
sizeUnit	-

resourceName	<i>A web tool for aligned text search</i>
resourceShortName	
downloadLocation	-
dateCreation	2012-
projectPartner	University of Belgrade, Faculty of Mathematics
iprHolder.organizationName	University of Belgrade, Faculty of Mathematics
contact.Person.surname	Zečević
contact.Person.givenName	Andelka
contact.Person.email	andjelkaz@matf.bg.ac.rs
DistributionInfo	underNegotiation
license	-
resourceLocation	-
distributionAccessMedium	accessibleThroughInterface
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	-
foreseenUse	human use
actualUse	human use
description	This is a web tool for effective search of aligned and annotated texts. It is especially designed for texts in which named entities were tagged. Its purpose is to compare annotation of NEs in aligned text and for that purpose a language independent classification schema for NEs is used.
relevantPublications	-
resourceType	technology tool / service
mediaType	text
lingualityType	multilingual
languageId	-
size	-
sizeUnit	-

resourceName	<i>Web applications (NE extraction from web pages)</i>
resourceShortName	
downloadLocation	-
dateCreation	2012-
projectPartner	University of Belgrade, Faculty of Mathematics
iprHolder.organizationName	University of Belgrade, Faculty of Mathematics
contact.Person.surname	Vitas
contact.Person.givenName	Duško
contact.Person.email	vitas@matf.bg.ac.rs
DistributionInfo	underNegotiation
license	-
resourceLocation	-
distributionAccessMedium	other -executable
restrictionsOfUse	academic-nonCommercialUse

licenseSignatory.Person.position	-
foreseenUse	NLP applications
actualUse	NLP applications
description	This is a web tool for extraction of proper names from categories given in Wikipedia for English, French, Serbian, Polish.
relevantPublications	-
resourceType	technology tool / service
mediaType	text
lingualityType	multilingual
languageId	EN/FR/SR/PL
size	-
sizeUnit	-

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

6.6. Slovak language resources of further interest – detailed specification

resourceTitle	<i>Slovak –Russian Parallel Corpus</i>
resourceName	sk-ru
urlDownload	http://korpus.juls.savba.sk/parus/
dateCreation	2006
projectPartner	LSIL
IPRholder.organizationShortName	LSIL/various
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
availability	available (pseudocorpus)
license	other
resourceLocation	LSIL
distributionMedium	internetBrowsing
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	-
actualUse.actualUse	human use; NLP
actualUse.useNLPspecific	-
description	The corpus contains original Russian fiction texts and their Slovak translations, with automatically aligned sentences.
relevantPublications	Garabík, Radovan, Захаров, Виктор Павлович: Параллельный русско-словацкий корпус. In: Труды международной конференции Корпусная лингвистика. Санкт-Петербург: Издательство С.-Петербургского университета 2006, p. 81 – 87.
urlDocumentation	http://korpus.juls.savba.sk/parus/
resourceType	corpus
resourceSubtype	parallel corpus
mediaType	text
noLanguages	2
multilingualityType	parallel
languageId	slk; rus
size	100 000
sizeUnit	sentence
annotationType	segmentation; lemmatization; PosTagging; MSD; alignment

resourceTitle	<i>Slovak –French Parallel Corpus</i>
resourceName	sk-fr

urlDownload	http://korpus.juls.savba.sk/frask/
dateCreation	2007
projectPartner	LSIL
IPRholder.organizationShortName	LSIL/various
contact.Person.surname	Garabík
contact.Person.givenName	Radovan
contact.Person.email	radovan.garabik@kassiopeia.juls.savba.sk
availability	available (pseudocorpus)
license	other
resourceLocation	LSIL
distributionMedium	internetBrowsing
restrictionsOfUse	academic-nonCommercialUse
licenseSignatory.Person.position	director
foreseenUse.foreseenUse	human use; NLP applications
foreseenUse.useNLPspecific	-
actualUse.actualUse	human use; NLP
actualUse.useNLPspecific	-
description	The corpus contains original French fiction texts and their Slovak translations, with automatically aligned sentences.
relevantPublications	VASILÍŠINOVÁ, Dorota, GARABÍK, Radovan: Parallel French-Slovak Corpus. In: Computer Treatment of Slavic and East European Languages. Proceedings of the conference Slovko 2007. Eds. J. Levická, R. Garabík. Brno: Tribun 2007.
urlDocumentation	http://korpus.juls.savba.sk/frask/
resourceType	corpus
resourceSubtype	parallel corpus
mediaType	text
noLanguages	2
multilingualityType	parallel
languageId	slk; fra
size	21 000
sizeUnit	sentence
annotationType	segmentation; lemmatization; PosTagging; MSD; alignment

resourceName	<i>Database of Root Morphemes</i>
resourceShortName	Database of root morphemes
downloadLocation	
dateCreation	2012
projectPartner	Prešov University
iprHolder.organizationName	Prešov University
contact.Person.surname	-
contact.Person.givenName	-

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

resourceName	<i>Dictionary of Slovak Adjective Collocations</i>
resourceShortName	Dictionary of Slovak Adjective Collocations
downloadLocation	
dateCreation	2012
projectPartner	LSIL
iprHolder.organizationName	University of St. Cyril and Methodius in Trnava
contact.Person.surname	-
contact.Person.givenName	-
contact.Person.email	-
DistributionInfo	underNegotiation
license	-
resourceLocation	University of St. Cyril and Methodius in Trnava
distributionAccessMedium	
restrictionsOfUse	-
licenseSignatory.Person.position	
foreseenUse	nlpApplications
actualUse	nlpApplications

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

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

size	440
sizeUnit	entry

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

7. Annex 2 – information for all identified language resources

7.1. Bulgarian

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	Linguali tyType	Outside the consortiu m	Publish ed in META- SHARE	Importan t to be published in META- SHARE	Describe why the resource is of further interest	Willing to publish resources in the META- SHARE	Describe the transformation work need to be done to make resources publishable
First D2.3												
Bulgarian National Corpus	BulNC	http://search.dcl.bas.bg	corpus	979000000	token	monolin gual	no	1,2,3	yes	The biggest general balanced corpus of Bulgarian	published	Published in 1st batch; enhanced in 2nd batch
Bulgarian PoS Annotated Corpus	BulPoSCor	http://search.dcl.bas.bg	corpus	150 000	word	multiling ual	no	1	yes	The only available POS annotated corpus	published	Published in 1st batch
Bulgarian Sense Annotated Corpus	BulSemCor	http://dcl.bas.bg/semcor/en/	corpus	105 000	word	monolin gual	no	1	yes	The only sense annotated corpus	published	Published in 1st batch
Bulgarian-X language parallel corpora	Bul-XCor	http://dcl.bas.bg/poscor/en/	corpus	19 000 000 000	token	monolin gual	no	1,2,3	yes	The biggest parallel corpus of Bulgarian	published	Published in 1st batch; enhanced in 2nd batch
Bulgarian wordnet	BulNet	http://catalog.elra.info/product_info.php-products_id=802	lexicalConceptu alResource	41000	synset	multiling ual	no	1,2,3	yes	The Bulgarian wordnet, 1/4th of the English	published	Published in 1st batch; enhanced in 2nd batch
Bulgarian FrameNet	BulFrameNet	http://dcl.bas.bg/LexIt	lexicalConceptu alResource	3 000	concept	monolin gual	no	no	yes	The only frame dictionary of Bulgarian	yes	Resource and software development, testing documentation.
Bulgarian Grammatical Dictionary	BulGram	http://dcl.bas.bg/est	lexicalConceptu alResource	85 000	lemma	monolin gual	no	no	yes	Based on the latest orthographic dictionary of Bulgarian	yes	Documentation

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	Linguali tyType	Outside the consortiu m	Publish ed in META- SHARE	Importan t to be published in META- SHARE	Describe why the resource is of further interest	Willing to publish resources in the META- SHARE	Describe the transformation work need to be done to make resources publishable
Second D2.3												
Corpus of Colloquial Bulgarian	BgSpeech	http://bgspeech.net/bg/resources/razg.html	corpus	534 604	word	monolin gual	yes	no	yes	The only available Corpus of Colloquial Bulgarian	will be published	META-SHARE compliance, data set enlargement
Diachronic corpus of Bulgarian Language	histdict	http://histdict.uni-sofia.bg	corpus	-	-	multiling ual	yes	3	yes	The only available Diachronic corpus of Bulgarian Language	yes	Negotiate license agreement.
Corpus of Spoken Bulgarian	SpokenBg	-	corpus	605 202	word	monolin gual	yes	3	yes	The only available Corpus of Spoken Bulgarian	will be published	META-SHARE compliance, data set enlargement
Bulgarian Spell checker	WinEst	http://dcl.bas.bg/sites/default/files/webfm/WinEst/winestSetup.exe	technologyTool Service	-	-	monolin gual	no	1	yes	Freely downloadable spell checker of Bulgarian	published	Published in 1st batch
Bulgarian Spell Checker Web Service	WebEst	http://dcl.bas.bg/est/index_en.php#tabs-5	technologyTool Service	-	-	monolin gual	no	1	yes	The only Bulgarian spell checker web service	published	Published in 1st batch
Chooser - annotation tool	Chooser	http://dcl.bas.bg/en/programs_en.html	technologyTool Service	-	-	monolin gual	no	2	yes	Multifunctional language independent annotation tool	published	Published in 2nd batch
Hydra - tool for developing wordnets	Hydra	http://dcl.bas.bg/en/programs_en.html	technologyTool Service	-	-	multiling ual	no	2	yes	Tool for editing, searching and validation wordnets	published	Published in 2nd batch
Bulgarian Sentence Splitter		http://dcl.bas.bg/en/programs_en.html	technologyTool Service	-	-	monolin gual	no	2	yes	Freely available	published	Published in 2nd batch

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LingualityType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
Bulgarian Tokenizer		http://dcl.bas.bg/en/programs_en.html	technologyTool Service	-	-	monolingual	no	2	yes	Freely available	published	Published in 2nd batch
TREFL – Translation Reference Library	TREFL	http://web.uniplovdiv.bg/rousni/index_fr.htm	lexical/conceptual resource; technology tool	1 GB	file	multilingual	yes	no	yes	Translation library	yes	Negotiate license agreement.
SARP- Speech Analyzer Rapid Plot. Plotting vowels in F2-F1 scatter charts with multiple data sets	SARP	http://web.uniplovdiv.bg/rousni/sarp	technology tool	170 MB	file	multilingual	yes	3	yes	Speech analyzer	will be published	Negotiate license agreement
RTComp - Real Time Comparison	RTComp	http://web.uniplovdiv.bg/rousni/rtcomp	technology tool	10 MB	file	multilingual	yes	no	yes		yes	Negotiate license agreement
Third D2.3												
Bulgarian-English Clause Aligned corpus	BulEnAC	http://dcl.bas.bg/en/corpora_en.html	corpus	30385	sentence	bilingual	no	3	yes	The only available clause aligned corpus of Bulgarian	yes	META-SHARE compliance
Lists of Bulgarian Multiword Expressions	BulMWEs	http://dcl.bas.bg/Resources/MWEs/lists.zip	lexicalConceptualResource	27784	multiword units	monolingual	no	2	yes	He only freely available lists of mwes of Bulgarian	published	Published in 2nd batch
Bulgarian Frequency Dictionary	BFD	http://dcl.bas.bg/Resources/Frequency/Frequency.zip	lexicalConceptualResource	27784	words	monolingual	no	2	yes	Freely available	published	Published in 2nd batch
ClauseAlign	ClauseAlign	http://dcl.bas.bg/en/programs_en.html	technologyTool Service	-	-	multilingual	no	3	yes	The only available clause aligning tool	yes	Tool development, META-SHARE compliance
BgMWE – tool for MWE and NE recognition	BgMWE	http://dcl.bas.bg/Tools/MWEs/bgMWE.jar	technologyTool Service	-	-	monolingual	no	2	yes	Freely available	published	Published in 2nd batch
Web based infrastructure for Bulgarian data processing	DCLservices	http://dcl.bas.bg/dclservices/registration/	technologyTool Service	-	-	monolingual	no	2	yes	Annotation web service	published	Published in 2nd batch

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LingualityType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
Bulgarian Word Sense Disambiguation Tool	BulWSD	http://dcl.bas.bg/en/programs_en.html	technologyTool Service	-	-	monolingual	no	no	yes	The only complex wsd tool	yes	Tool development, META-SHARE compliance
Bulgarian Spell Checker for Mac	MacEst	http://dcl.bas.bg/en/MacEst-en.html	technologyTool Service	-	-	monolingual	no	3	yes	Freely available	will be published	Tool development, META-SHARE compliance
Bulgarian Grammar Checker for Windows	WinEst+	http://dcl.bas.bg/est/index_en.php#tabs-5	technologyTool Service	-	-	monolingual	no	3	yes	Freely available	will be published	Tool development, META-SHARE compliance
Bulgarian Grammar Checker Web Service	WebEst+	http://dcl.bas.bg/est/index_en.php#tabs-5	technologyTool Service	-	-	monolingual	no	3	yes	Web service	will be published	Tool development, META-SHARE compliance

7.2. Croatian

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LingualityType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
First D2.3												
Croatian National Corpus	HNK	http://hmk.ffzg.hr	corpus	101 884 284	token	monolingual	no	1	yes	The first over 100 M representative corpus of Croatian.	yes	Tool development, META-SHARE compliance
Croatian Morphological Lexicon	HML	http://hmk.ffzg.hr	lexicalConceptualResource	4 000 000	entry	monolingual	no	1	yes	The largest Croatian Morphological Lexicon.	yes	Tool development, META-SHARE compliance
Croatian-English Parallel Corpus	Hr-En p-corp	http://hmk.ffzg.hr/hr-en_p-corp	corpus	62 566	translation unit	bilingual	no	1	yes	A parallel unidirectional (hr to en) corpus of contemporary Croatian standard language collected from articles appearing in Croatia Weekly newspapers, published from 1998 to 2000.	yes	Published in batch 1

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
Croatian Lemmatisation Server	CLS	http://hml.ffzg.hr	technology-Tool	-	-	monolingual	no	1	yes	A web-based service for lemmatisation, POS- and MSD-tagging of Croatian texts. It accepts input in two modes: web HTML form and upload mode.	yes	Tool development, META-SHARE compliance
Croatian Valency Dictionary	CROVALLEX	http://cal.ffzg.hr/crovallex/index.html	lexicalConceptualResource	1740	entry	monolingual	no	1	yes	The Croatian Valency Lexicon of Verbs, Version 2.0008 (CROVALLEX 2.0008) is an attempt of formal description of valency frames of Croatian verbs.	yes	Enlarging the CROVALLEX with new entries.
Second D2.3												
Croatian Web Corpus	hrWaC	http://www.nljub.esic.net/resources/corpora/hrwac/	corpus	1 186 795 086	token	monolingual	no	2	yes	Croatian Web Corpus (hrWaC) is the largest collected corpus for Croatian so far.	yes	Encoding as TEI P5
Slovene Web Corpus	slWaC	http://www.nljub.esic.net/resources/corpora/slzac/	corpus	380 299 844	token	monolingual	no	2	yes	The first Slovenian web corpus. It was collected together with Croatian web corpus.	yes	Lemmatisation and MSD-tagging.
Croatian-English Parallel Web Corpus	hr-enWaC	http://www.nljub.esic.net/resources/corpora/hrenwac/	corpus	4 397 887	token	multilingual	no	2	yes	Enhancement of hr-en parallel corpus.	yes	Sentence segmentation and alignment.
Croatian Dependency Treebank	HOBS	http://hobs.ffzg.hr/download	corpus	4 500	sentence	monolingual	no	2	yes	The first Croatian treebank.	yes	Conversion into conll format
SouthEast European Parallel Corpus	SETimes Corpus	http://www.nljub.esic.net/resources/corpora/setimes/	corpus	43 142 458	token	multilingual	no	2	yes	The parallel corpus in 10 SE European languages that gives the unique opportunity to investigate the language similarities and differences between them.	yes	Sentence segmentation and alignment.
Web Content Extractor	WebContentExtractor	http://www.nljub.esic.net/resources/tools/webcontentextractor/	technology-Tool	-	-	multilingual	no	2	yes	Algorithm for web crawling and boilerplate removal in web pages for very large web-based corpora	yes	Licensing mechanism

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
Collocation and Term Extractor	CollTerm	http://www.nljubjesic.net/resources/tools/collterm/	technology-Tool	-	-	multilingual	no	2	yes	Language independent collocation and term extractor. Calibrated for Croatian	yes	Evaluation and calibration for Croatian
Third D2.3												
Croatian Wordnet	CroWN	http://hnk.ffzg.hr/crown	lexicalConceptualResource	ca 10 000	synset	monolingual	no	3	yes	The first Croatian semantic network describing the relations between words with established methodology.	yes	Enhancing the crown to at least 10,000 synsets.
Croatian Translations of Acquis	hrAcquis	http://hnk.ffzg.hr/hracquis	corpus	60 000 000	token	multilingual	no	3	yes	The integral translations of Acquis Communautaire to Croatian. It will be aligned with other language translations of Acquis.	yes	Conversion to XML according the JRC-Acquis DTD, sentence alignment with all EU official languages.
Corpus of Narodne novine	NN-corp	http://hnk.ffzg.hr/nn	corpus	15 000 000	token	monolingual	no	3	yes	The corpus of texts from the Official Journal of the Republic of Croatia.	yes	Crawling, sentence segmentation, lemmatisation, MSD-tagging.
Croatian Language Corpus	Riznica	http://riznica.ihj.hr	corpus	70 000 000	token	monolingual	yes	3	yes	The largest Croatian diachronic corpus.	yes	Metadata description.
Croatian Lemmatisation Web Service	CroLem	http://lt.ffzg.hr/crolem	technologyToolService			monolingual	no	3	yes	The enhancement of Croatian Lemmatisation Server that features disambiguation and enlarged Croatian Morphological Lexicon behind it.	yes	Programming the interface.
Croatian NERC Web Service	CroNERC	http://lt.ffzg.hr/cronerc	technologyToolService			monolingual	no	3	yes	The first Croatian NERC as a web service.	yes	Programmin the interface.
Croatian Speech Corpus	CroSpeak Corpus	http://www.inf.uniri.hr/~ivoi/CROSPEECH/index.htm	corpus	227 280	token	monolingual	yes	no	yes	The first Croatian speech corpus.	yes	Metadata description.
Croatian Academic Spelling Checker	Hascheck	http://hacheck.tel.fer.hr/	technologyToolService			monolingual	yes	no	yes	The oldest Croatian spelling checker, functioning on-line since 1994. Has been applied to more than 100 Mw of Croatian texts and its collected database of Croatian types represents the valuable language resource.	yes	Standardisation of web service access (REST protocol).

7.3. Hungarian

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
First D2.3												
Hunglish Parallel Corpus	Hunglish	BME MOKK	corpus	54200000	token	bilingual	yes	1	yes	This is the only Hungarian-English parallel corpus which is under constant maintenance and is the most visited parallel corpus of the Hungarian language.	yes	Tool development, META-SHARE compliance
Hungarian Wordnet	HuWN	BME MOKK	lexical conceptual resource	42000	synset	monolingual	no	1	yes	The only Hungarian part of the wordnet ontology.	yes	Tool development, META-SHARE compliance Negotiate license agreement
Szeged NER Corpus		Szeged University	corpus	220000	token	monolingual	yes	1	yes	Szeged NER corpus is the most precise NER tagged corpus of the Hungarian language. It was built manually as a part of the Szeged Treebank.	yes	Tool development, META-SHARE compliance
Szeged Corpus		Szeged University	corpus	1200000	token	monolingual	yes	1	yes	A morpho-syntactically annotated and manually disambiguated corpus of 1,2 million words.	yes	Tool development, META-SHARE compliance Negotiate license agreement
Szeged Treebank		Szeged University	corpus	1200000	token	monolingual	yes	1	yes	A manually checked treebank of 1,2 million words.	yes	Tool development, META-SHARE compliance Negotiate license agreement
Hungarian Webcorpus		Szeged University	corpus	1,48E+09	token	monolingual	yes	1	yes	At this time this is the biggest webcorpus of the Hungarian language.	yes	Tool development, META-SHARE compliance Negotiate license agreement
morphdb.hu	morphdb.hu	BME MOKK	lexical / conceptual resource	400 000	item	monolingual	yes	1	yes	A formal morphological description for Hungarian is intended to be used by the hunmorph morphological analyzer	yes	Improvements to the documentation

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LingualityType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
hunmorph	hunmorph	http://mokk.bme.hu/resources/hunmorph	technology tool /service	-	-	monolingual	yes	1	yes	Hungarian lexical database and morphological grammar, used in most of morphological applications for Hungarian language.	yes	Metadata collection, testing, Minor improvements to the documentation
hunalign	hunalign	http://mokk.bme.hu/resources/hunalign	technology tool / service	-	-	bilingual	yes	1	yes	A sentence aligner for several languages. 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.	yes	Tool development, META-SHARE compliance Negotiate license agreement
huntoken	huntoken	http://mokk.bme.hu/resources/huntoken	technology tool / service	-	-	monolingual	yes	1	yes	The best tokenizer for Hungarian language	yes	Writing a wrapper around huntoken that integrates it into the hunalign-harness parallel text processing pipeline, Testing, Minor edits to the documentation
BABELHungarian Clear Speech Database	BABEL	hosted in international repository (ELRA)	corpus	4	hour	monolingual	no	1	yes	Audio corpus allowing for research on speech analysis, speech syntax, speech prosody and phonetics in general.	yes	Metadata annotation, extension, enhancement
Hungarian Reference Speech Database	MRBA	BME-TMIT and University of Szeged	corpus	6,5	hour	monolingual	yes (partly)	2	yes	Audio corpus designed to train broadband ASR models.	yes	Documentation improvement, Metadata annotation, IPR clearance, Phoneme level segmentation extension and checking of segmentations
Hungarian Telephone Speech Database	MTBA	BME-TMIT and University of Szeged	corpus	5	hour	monolingual	yes (partly)	2	yes	Audio corpus designed to train telephone band ASR models.	yes	Documentation improvement, Metadata annotation, IPR clearance, Phoneme level segmentation extension and checking of segmentations
Hungarian Telephone Client Speech Database	MTÜBA	hosted by owner	corpus	60	hour	monolingual	no	2	yes	Semi-spontaneous telephone environment speech corpus.	yes	Tool development, META-SHARE compliance Negotiate license agreement

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
Broadcast News Database	BND	BME-TMIT	corpus	5	hour	monolingual	no	1	yes	Audio-video multimodal speech database for research. Has its counterparts for a variety of languages.	yes	Database cleanup, upgrade for interoperability (transcriptions and audio), documentation harmonization, clearance of license terms, metadata description
Emotion Database	-	BME-TMIT	corpus	50	hour	monolingual	no	1	yes	The only available spontaneous speech emotion corpus for Hungarian.	yes	Partly newly created database, restructurisation, documentation, transcription, extension, metadata annotation, IPR clearance
Sound Gesture Database	SGD	BME-TMIT	lexical conceptual resource	770	token	monolingual	no	1	yes	The only available spontaneous speech emotion lexicon for Hungarian.	yes	Partly newly created lexicon, Structurisation, documentation, extension, metadata annotation, IPR clearance
Medical Database	MeD	BME-TMIT and Semmelweis University	corpus	1	hour	monolingual	yes (partly)	3	yes	The only available pathological speech corpus for Hungarian.	yes	Newly created database, Structurisation, documentation, segmentation, metadata annotation, IPR clearance
Broadcast Lectures Database	BLD or ME corpus	ME	corpus	~150+	hour	monolingual	no	3	yes	Very wide lecture topics presented orally by various scientist.	yes	Metadata annotation, IPR clearance, extension, standardization
Hungarian Speech Database of Holocaust Survivors' Testimonies	Hungarian MALACH	AITIA International	corpus	31	hour	monolingual	yes	3	yes	The only available spontaneous elderly speech corpus for Hungarian.	yes	Metadata annotation, IPR clearance, anonymization (waveform and text)
Hungarian Parliamentary Speeches	HuPaS	BME-TMIT	corpus	1000+	hour	monolingual	no	2	yes	One of the biggest publicly available Hungarian speech corpus	yes	Metadata annotation, selection of accurately aligned text and speech segments via Large Vocabulary Continuous Speech Recognition (LVCSR) technology
Word level speech database	Words-hu	BME-TMIT	corpus	3681	seconds	monolingual	yes	1	yes	It contains all the Hungarian sound combinations.	yes	Cleanup, extension, restructurisation, metadata annotation and IPR clearance
Second D2.3												

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
Hungarian NER Corpus based on Wikipedia	hunNERwiki	BME MOKK	corpus	ca. 19 million	token	monolingual	yes	2	yes	The largest ever NE-tagged corpus for Hungarian, which can be used for training and testing NE recognizer applications; comparable with the performance of other state-of-the-art systems.	yes	Tool development, META-SHARE compliance Negotiate license agreement
Hungarian Opinion-Tagged Sentence Bank	OpinHuBank	GeoX Kft.	corpus	10 000	annotated sentence			2	yes	The biggest human-annotated resource for researching, evaluating and developing opinion mining systems for Hungarian language.	yes	Upgrade of the NER tools, annotation of the polarity of each sentence, enhancement of the documentation
Hungarian WSD Corpus	HuWSD	Szeged University	corpus	300-500x39	text	monolingual	yes	2	yes	The only corpus for Hungarian made for word sense disambiguation.	yes	The website of the corpus has been updated, A few XML errors have been corrected, The documentation of the resource has been improved
Szeged Criminal NE Corpus	SzegedCrimNE	Szeged University	corpus	540K	token	monolingual	yes	2	yes	Corpus for Hungarian language of annotated named entities. Furthermore the corpus is made in two versions one contains tag-for-tag annotation, the other contains tag-for-meaning annotation.	yes	The website of the corpus has been updated, a few annotation errors have been corrected
Szeged Treebank FX	Szeged Treebank FX	Szeged University	corpus	82K	sentence	monolingual	yes	2	yes	The Szeged Treebank is a unique corpus of manually annotated light verb constructions (contains 6734 occurrences of 1215 light verb constructions).	yes	Tool development, META-SHARE compliance
SzegedParalell	SzegedParalell	Szeged University	corpus	99K	sentence alignment units	bilingual	yes	2	yes	The English-Hungarian parallel corpus contains texts selected on the basis of grammatical and translational criteria. Both paragraph and sentence alignment were checked and corrected manually.	yes	The website of the corpus has been updated, A few alignment errors have been corrected, The documentation of the resource has been improved

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
SzegedParallelFX	SzegedParallelFX	Szeged University	corpus	14K	sentence alignment units	bilingual	yes	2	yes	Semi-compositional units annotated in a parallel corpus	yes	Tool development, META-SHARE compliance
Hungarian Verb Phrase Constructions	HVPC	RIL HAS	lexical / conceptual resource	6 200	unit	monolingual	no	2	yes	Unique resource of the Hungarian verb phrase constructions (are extracted from the Hungarian National Corpus).	yes	Conversion of the original database to standard representation using LMF (Lexical Markup Framework)
Hungarian Language Processing Tools in NooJ	NooJ	http://corpus.nyttud.hu/nooj	lexical / conceptual resource, technology tool / Service	~10	file	monolingual	no	2	yes	Unique resources for a broad scale of morphological, syntactic, lexical, semantic and psychological content analyses.	yes	Upgrade of the main dictionaries
hunner	hunner	http://mokka.bme.hu/resources/huntag	technology tool / service	-	-	monolingual	yes	2	yes	Hunner is the most frequently used NER-tagger for Hungarian language analysis	yes	Tool development, META-SHARE compliance
hunpars	hunpars	http://mokka.bme.hu/resources/hunpars	technology tool / service	-	-	monolingual	yes	2	yes	Hunpars is the best syntactic analyzer developed for Hungarian language	yes	Metadata collection, minor bugfixes, refreshing of the English documentation, documentation enhancement
hunpos	hunpos	http://mokka.bme.hu/resources/hunpos	technology tool / service	-	-	monolingual	yes	2	yes	Hunpos is an open source reimplementation of tnt made for the Hungarian language	yes	Metadata collection, minor bugfixes, refreshing of the English documentation, documentation enhancement
Hungarian read Poem, "Sir John" aligned text	HunReP-SJ-text	THINKTech Nonprofit Ltd.	corpus	5k	word	monolingual	yes	2	no	The only publicly available Hungarian language aligned read poem text	yes	Tool development, META-SHARE compliance
Hungarian AudioBook, "Eclipse of the Crescent Moon" aligned text	HunAuB-ECM-Text	THINKTech Nonprofit Ltd.	corpus	30k	word	monolingual	yes	2	no	The only publicly available Hungarian language aligned audiobook text	yes	Tool development, META-SHARE compliance
Automatic Prosodic Segmenter	ProSeg	BME-TMIT	technology tool / service	-	other	multilingual	no	3	yes	Unique multilingual prosodic segmentation tool	yes	Documentation, bug-fix, standardisation, metadata, IPR clearance
Hungarian Phonetic Transcriber	HunPhoner	BME-TMIT	technology tool / service	-	other	monolingual	no	3	yes	The only available phonetic transcriber for Hungarian	yes	Documentation, bug-fix, standardisation (character encoding options), metadata, IPR clearance

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
Third D2.3												
Hungarian National Corpus	HNC	corpus.nyttud.hu/hnc	corpus	187600000	token	monolingual	no	3	yes	The national corpus of Hungarian derived into five subcorpora by regional language variants, and into five subcorpora by text genres also.	yes (through a query interface)	Extensive upgrade on tokens, upgrade on analyzer and metadata.
Hungarian Spontaneous Speech Database	BEA	HASRIL	corpus	270	hours	monolingual	no	3	yes	BEA is a multi-functional speech database of Hungarian that contains various types of spontaneous speech (including conversations) and sentence repetitions and reading. This is the largest speech database of Hungarian consisting of about 270 hour recorded speech material of 265 speakers at present.	yes	Increase the number of speakers (up to 500), increase the number of annotated materials by speech sound level incorporate natural language processing technology.
Child language corpus	CHILC	HASRIL	corpus	60	interview	monolingual	no	3	no	The child language corpus of interviews including several tasks (picture-based story-telling, telling the rules of a well-known game) and guided conversation. The resource is available in chat (CHILDES) transcription format.	no	Prepare the xml tagged version with lemmatization, morphological analysis and POS tagging.
Hungarian Human-Computer Interaction Technologies Multimodal Database	HUCOMTECH	University of Debrecen	multimodal corpus	50	hours	monolingual	yes	3	yes	The first and only resource for Hungarian language of aligned text-video-audio segments. The alignment is made by speech units.	yes	Database format conversion, database enhancement, corrections of the aligned speech units.
ht-online	ht-online	Termini Research Network	lexical / conceptual resource	4000	entry	monolingual	yes	3	no	A unique lexical database of the most common loanwords in Hungarian language used outside Hungary (collected from 7 regions).	no	Xml conversion, TEI compatibility, English documentation, upgrade with more headwords.

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
Hungarian concise dictionary (with sample sentences)	HCD	TINTA Publishing House	lexical / conceptual resource	16000	entry	monolingual	yes	3	yes	A unique dictionary of Hungarian language of 16 000 headwords (entries) followed by frequency data.	no	Xml conversion, TEI compatibility, correction of discrepancies in the structure of entries.
High-speed Unification Morphology	HUMor	MorphoLogic Ltd.	tool	100000	entries	monolingual	yes	3	no	The system is language-independent, allowing multilingual applications for a variety of language types	no	Bugfixing, programming tasks.
Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection	ParallelSpeech-hu	BME-TMIT	corpus	25,5	hours	monolingual	no	2	yes	This speech database contains 2000 sentences. Each speaker read this sentence set. This parallel speech database is used to train HMM based TTS and for unit selection TTS.	yes	Tool development, META-SHARE compliance
Graphical query interface for Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection	-	University of Debrecen	technology tool / service	-	other	monolingual	yes	3	yes	An important tool simplifying access to the corpus and allowing for easy and fast statistical processing.	yes	Newly created service, creation, documentation, metadata, IPR.
Read Speech Database for Hungarian	ReadSpeech-hu	BME-TMIT	corpus	10	hours	monolingual	no	2	yes	The read speech database contains sentences from weather forecast news. The sentence collection represents the four seasons. This database can be used for analysing speech characteristics in weather forecast news and also as the basic speech database of a corpus based Concept-to-Speech system.	yes	Tool development, META-SHARE compliance
Hungarian BABEL phonetic and prosodic segmentation and syntactic analysis	BABEL-Addon1	BME-TMIT	corpus	330	utterances	monolingual	no	2	yes	Audio corpus allowing for research on speech analysis, speech syntax, speech prosody and phonetics in general.	yes	Tool development, META-SHARE compliance

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
Di-phone database for text-to-speech conversion in Hungarian	Di-phone-hu	BME-TMIT	corpus	1646	seconds	monolingual	no	2	yes	Contains combinations of 38 sounds for TTS conversion; can be used for educational purposes and in speech research.	yes	Tool development, META-SHARE compliance
Tesztel Hungarian Noisy Telephone Speech Corpus	-	BME-TMIT	corpus	100	speakers	monolingual	no	no	no	Noisy speech samples for noise robust ASR in Hungarian	no	-
A Hungarian Child Database for Speech Processing Applications	-	BME-TMIT	corpus	72	speakers	monolingual	no	no	no	Child speech for Hungarian, useful for speech training and language learning applications for children.	no	-
Multilingual speech segmentation tool	-	BME-TMIT	tool	-	-	multilingual	no	no	no	Multilingual speech segmentation tool.	no	-
Sentence modality recognizer	-	BME-TMIT	tool	-	-	bilingual	no	no	no	Sentence modality recognizer from speech for Hungarian and German, can be used in speech recognition and understanding.	no	-
BABEL and MRBA sentence modality annotation for Hungarian	-	BME-TMIT	corpus	50	speakers	monolingual	no	no	no	Corpus holding modality annotations for subparts of Hungarian BABEL and MRBA corpora.	no	-

7.4. Polish

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
First D2.3												
Polish Sejm Corpus	PSC	http://clip.ipipan.waw.pl/PSC	corpus	114000000	token	monolingual	no	1, 2	yes	Specialized quasi-spoken corpus, preparing ground to be aligned with audio/video data	yes	Data conversion, IPR status clarification, data set enlargement in subsequent batches
PoliMorf morphological dictionary	PoliMorf	http://zil.ipipan.waw.pl/PoliMorf	lexicalConceptualResource	6382227	entry	monolingual	yes/no	1, 2	yes	Largest morphological dictionary of Polish	yes	Data merger and manual correction of entry, additional classification, IPR clarification
Polish WordNet	plWordNet	http://plwordnet.pwr.wroc.pl/wordnet	lexicalConceptualResource	73000	synset	bilingual	yes	1, 2	yes	Largest wordnet of Polish, one of the largest in the world	yes	META-SHARE compliance, data set enlargement
1 million subcorpus of National Corpus of Polish	1MNKJP	http://clip.ipipan.waw.pl/LRT-action=AttachFile&amp;do=get&amp;target=NKJP-PodkorpusMilionowy-1.0.tgz	corpus	1003956	word	monolingual	yes/no	1	yes	Manually annotated subcorpus National Corpus of Polish, multiple annotation levels	yes	META-SHARE compliance
Polish Named Entity Gazetteer	PNEG	http://clip.ipipan.waw.pl/Gazetteer	lexicalConceptualResource	44944	entry	monolingual	yes/no	1, 2	yes	Large gazetteer of Polish	yes	Legal status clarification, preparation of the LMF version
LUNA.PL Corpus	LUNA.PL	http://zil.ipipan.waw.pl/LUNA	corpus	81049	word	monolingual	no	1	yes	Manually transcribed spoken data	yes	Data conversion, META-SHARE compliance
LUNA-WOZ.PL Corpus	LUNA-WOZ.PL	http://zil.ipipan.waw.pl/LUNA	corpus	5523	utterance	monolingual	no	1	yes	Manually transcribed spoken data	yes	Data conversion, META-SHARE compliance
Polish Parallel Corpora	PPC	http://www.clip.waw.pl/Polish_Parallel_Corpora	corpus	2000000	word	multilingual	no	1	yes	-	yes	-

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work needed to be done to make resources publishable
Polish Spoken Multimedia Corpus	SMC	http://www.clip.waw.pl/Polish_Spoken_Multimedia_Corpus	corpus	2500000	word	monolingual	no	2	yes	-	yes	-
Polish Spoken Conversational Corpus	PSCC	http://www.nkjp.uni.lodz.pl/spoken.jsp	corpus	1500000	word	monolingual	no	1	yes	-	yes	-
Second D2.3												
Morphosyntactic tagset converter for positional tagsets	TaCo	http://zil.ipipan.waw.pl/TaCo	toolService	-	-	monolingual	no	2	yes	Useful due to long-lasting problems with tagset conversion between Polish NLP tools	yes	Tool development and standardization, META-SHARE compliance
Spejd	Spejd	http://zil.ipipan.waw.pl/Spejd	toolService	-	-	multilingual	no	2	yes	Successfully used in other projects and for other languages	yes	META-SHARE compliance, documentation improvements
N-grams from balanced National Corpus of Polish	NKJPNGrams	http://zil.ipipan.waw.pl/NKJPNGrams	corpus	5364398	unigram	monolingual	no	2	yes	Successfully used for language modelling in other projects	yes	Creation of the resource based on NKJP, IPR clarification
Distributable subcorpus of National Corpus of Polish	DistrNKJP	http://zil.ipipan.waw.pl/DistrNKJP	corpus	99280766	word	monolingual	no	2	yes	Redistributable part of the largest national corpus	yes	Creation of the resource based on NKJP, IPR clarification
Morfeusz morphological analyzer	Morfeusz	http://sgjp.pl/morfeusz/index.html.en	toolService	-	-	monolingual	yes/no	2	yes	Successfully used by numerous scientific projects	yes	Update of the tool using polymorf data, IPR clarification
Morfologik Inflectional Dictionary	Morfologik	http://sourceforge.net/projects/morfologik	toolService	-	-	monolingual	yes/no	2	yes	Successfully used by numerous scientific and open-source projects, e.g. Openoffice.org	yes	Update of the data representation using polymorf data
Grammatical Lexicon of Polish Phraseology	SEJF	http://zil.ipipan.waw.pl/SEJF	lexicalConceptualResource	3176	entry	monolingual	yes/no	2	yes	Valuable resource of mwus	yes	Data conversion, IPR clarification

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work needed to be done to make resources publishable
Grammatical Lexicon of Polish Economical Phraseology	SEJFEK	http://zil.ipipan.waw.pl/SEJFEK	lexicalConceptualResource	11212	entry	monolingual	yes/no	2	yes	Valuable resource of domain mwus	yes	Data conversion, IPR clarification
Grammatical Lexicon of Warsaw Urban Proper Names	SAWA	http://zil.ipipan.waw.pl/SAWA	lexicalConceptualResource	9000	entry	monolingual	yes/no	2	yes	Valuable resource of domain mwus	yes	Data conversion, IPR clarification
Multilingual lexicon of toponyms	WikiTopoPl	http://zil.ipipan.waw.pl/WikiTopoPl	lexicalConceptualResource	155000	entry	multilingual	yes/no	2	yes	Wikipedia-based, successfully used by other projects	yes	Data conversion, META-SHARE compliance
Polish Valence Dictionary	Walenty	http://clip.ipipan.waw.pl/Walenty	lexicalConceptualResource	1438	entry	monolingual	yes/no	2	yes	Aiming towards the largest valence dictionary of Polish	yes	IPR clarification, data merging, manual correction
Summarizer	Summarizer	http://clip.ipipan.waw.pl/Summarizer	toolService	-	-	monolingual	yes	2	yes	One of the few summarizers for Polish	yes	IPR clarification, META-SHARE compliance
Morfologik-stemming	Morfologik-stemming	http://sourceforge.net/projects/morfologik/	toolService	-	-	monolingual	yes	2	yes	Successfully used by numerous scientific and open-source projects	yes	Update of the tool using polimorf data
Corpus of the Polish language of the 1960s	PL196x	http://clip.ipipan.waw.pl/PL196x	corpus	500000	word	monolingual	yes/no	2	yes	Used by numerous projects, e.g. Speech recognition training, tagger training	yes	Manual correction of data, IPR clarification
Shallow Grammar for the National Corpus of Polish	NKJPgrammar	http://zil.ipipan.waw.pl/ShallowGrammars	lexicalConceptualResource	1187	rules	monolingual	no	2	yes	Successfully used in preparation of the National Corpus of Polish and in numerous other applications (e.g. NP detection)	yes	Grammar improvement, META-SHARE compliance
Pantera	Pantera	http://zil.ipipan.waw.pl/Pantera	toolService	-	-	monolingual	no	2	yes	Best Brill tagger for Polish	yes	Implementation and documentation improvements
PolNet	PolNet	http://lrc.amu.edu.pl/polnet/index.php	lexicalConceptualResource	13200	synset	monolingual	yes	2	yes	Second largest wordnet of Polish	yes	IPR clarification, META-SHARE compliance
Polish-Russian Parallel Corpus	PolRosPC	-	corpus	25 000 000	word	bilingual	yes	3	yes	-	yes	-

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work needed to be done to make resources publishable
Polish Radio Żak and Radio Łódź Speech Corpus	RadioZakŁódź	http://www.zak.lodz.pl/ , http://www.radiolodz.pl/	corpus	50 000	word	monolingual	yes	no	no	-	no	-
Dictionary Of Selected English Collocations	DOSEC	-	lexical / conceptual resource	1 609 152	entry	monolingual	no	3	yes	-	yes	-
Dictionary of Selected Polish Collocations	DoSPiC	-	lexical / conceptual resource	2 500 000	entry	monolingual	no	3	yes	-	yes	-
Third D2.3												
LFG Grammar of Polish	LFGGrammarPL	-	lexicalConceptualResource	yet unknown	entry	monolingual	no	3	yes	Best LFG grammar for Polish	yes	Implementation and documentation improvements
Lexeme Forge	LexemeForge	-	toolService	-	-	monolingual	no	3	yes	Advanced tool for development of morphological dictionaries	yes	Implementation and documentation improvements
Slowal	Slowal	-	toolService	-	-	monolingual	no	3	yes	Advanced tool for development of valence dictionaries	yes	Implementation and documentation improvements
Lakon	Lakon	-	toolService	-	-	monolingual	yes	3	yes	One of the few summarizers for Polish	yes	IPR clarification, META-SHARE compliance
Składnica	Składnica	-	corpus	8227	sentence	monolingual	yes/no	3	yes	Largest treebank of Polish	yes	IPR clarification, META-SHARE compliance
Świgr	Świgr	-	toolService	-	-	monolingual	yes/no	no	yes	Deep parser of Polish	yes	IPR clarification, META-SHARE compliance
Formal Grammar of Polish	GFJP	-	lexicalConceptualResource	460	rules	monolingual	yes/no	no	yes	Large grammar for deep parsing of Polish	yes	IPR clarification, META-SHARE compliance
Anotatornia	Anotatornia	-	toolService	-	-	monolingual	yes/no	no	yes	Advanced annotation environment	yes	Implementation and documentation improvements
Ruler	Ruler	-	toolService	-	-	monolingual	yes/no	no	yes	Coreference resolver for Polish	yes	Tool documentation

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work needed to be done to make resources publishable
The corpus of Polish summaries	SummaryCorpus	–	corpus	yet unknown	text	monolingual	yes/no	3	yes	Large corpus of manually created extractive and abstractive summaries	yes	IPR clarification, format conversion
The parallel English-Polish corpus	ParallelCorpus	–	corpus	3000000	word each side	bilingual	yes/no	3	yes	Large, manually sentence-aligned Polish-English corpus with texts coming from several domains	yes	Manual alignment, format conversion
Syntactic-generative dictionary of Polish verbs	SSGCP	–	lexicalConceptualResource	10559	verb entry	monolingual	yes	3	yes	Large frame/valence dictionary of Polish	yes	IPR clarification, data conversion
PolSumm	PolSumm	–	toolService	–	–	monolingual	yes	no	yes	One of the few summarizers for Polish	yes	IPR clarification, META-SHARE compliance
Redistributable Polish-Russian Corpus	DistrPLRU	–	corpus	yet unknown	word	bilingual	yes/no	3	yes	Large, redistributable Slavic resource	yes	Data alignment, format standardization
Polish OpenCYC lexicon	OpenCYCPL	–	lexicalConceptualResource	yet unknown	word	bilingual	yes/no	3	yes	Advanced, manually created multilingual ontology	yes	Translation, alignment, format standardization
Voicelab SNUV Speech database	SNUV	snuv.pl	speech database	200	hour	monolingual	yes	3	yes	When acquired this will be the biggest freely available speech database of this sort	yes	>90% of the data acquired within CESAR in a massive crowdsourcing effort. Conversion, documentation, IPR clearance.
PELCRA Time-Aligned Spoken Corpus	TASC	pelcra.pl/tasc	speech database	40	hour	monolingual		2	yes	Unique time-aligned corpus of conversational Polish	yes	Manual time-alignment, conversion, TEI-encoding.
Polish-English Wikipedia NE dictionaries	NERDict	pelcra.pl/nerdict	lexical / conceptual resource		entry	bilingual	no	3	yes	-	yes	Export, cross-language linking, refinement.
Paralela DB	Paralela	pelcra.pl/paralela	corpus	50M	word	multilingual	yes/no	3	yes	A multilingual parallel corpus covering all the CESAR languages	yes	Web-crawling, alignment, conversion to TEI
Voicelab speech recognition engine	VoicelabEngine	voicelab.pl	tool	-	-	monolingual	yes	3	yes	Speech recognition platform for Polish	yes	META-SHARE compliance

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
Language detector	LDetect	pelcra.pl	tool	-	-	multilingual	no	3	yes	Allows distinguishing Polish from other languages in a multilingual text indexing scenario.	yes	Resource and software development, testing documentation.
Polish-English Spoken Learner Corpus	PESLC	-	speech database	50 000	word	bilingual	no	3	yes	Bilingual speech corpus of Poles switching between Polish and English.	yes	Transcription, time-alignment

7.5. Serbian

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
First D2.3												
Corpus of Contemporary Serbian	SrpKor	korpus.matf.bg.ac.rs	corpus	113000000	word	monolingual	no	1, 2	yes	The reference corpus of contemporary written Serbian	yes	Published in 1st batch; enhanced in 2nd batch
French-Serbian Aligned Corpus	SrpFranKor	korpus.matf.bg.ac.rs	corpus	1700000	word	bilingual	no	1, 2	yes	The only publicly available aligned French/Serbian corpus	yes	Published in 1st batch; enhanced in 2nd batch
English-Serbian Aligned Corpus	SrpEngKor	korpus.matf.bg.ac.rs	corpus	2000000	word	bilingual	no	2	yes	The only publicly available aligned English/Serbian corpus	yes	Published in 2nd batch
Serbian Lemmatized and PoS Annotated Corpus	SrpLemKor	korpus.matf.bg.ac.rs	corpus	3763352	word	monolingual	no	1	yes	The only publicly available pos and lemmatized corpus of Serbian	yes	Published in 1st batch
Multilingual Edition of Verne's Novel "Around the World in 80 Days"	Verne80days	korpus.matf.bg.ac.rs	corpus	1215839	word	multilingual	no	1, 2	yes	Multilingual text in 18 languages	yes	Published in 1st batch; enhanced in 2nd batch
Serbian Wordnet	SrpWN	korpus.matf.bg.ac.rs	lexicalConceptualResource	17550	synset	multilingual	no	1, 2	yes	Basic ontology database	yes	Published in 1st batch; enhanced in 2nd batch
Serbian Morphological Dictionary (Multext)	SrpMD	korpus.matf.bg.ac.rs	lexicalConceptualResource	85721	lemma	monolingual	no	2	yes	Comprehensive morphological e-dictionary of Serbian	yes	Published in 2nd batch
Serbian Named Entity Resources	SrpNER	korpus.matf.bg.ac.rs	lexicalConceptualResource	35000	lemma	monolingual	no	no (3)	yes	Comprehensive morphological e-dictionary of Serbian proper names	yes	Negotiate license agreement
AlfaNum Morphologic Dictionary of Serbian	AlfaNum MD	alfanum.ftn.uns.ac.rs	lexicalConceptualResource	100517	lemma	monolingual	yes	no (3)	yes	The morphological dictionary of Serbian with accentuation	no	-

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
AlfaNum Text Corpus of Serbian	AlfaNumKor	alfanum.ftn.uns.ac.rs	corpus	200027	word	monolingual	yes	no (3)	yes	For TTS and ASR applications	no	-
AlfaNum Speech Databases for ASR	AlfaNum ASR	alfanum.ftn.uns.ac.rs	corpus	-	-	monolingual	yes	no (3)	yes	For TTS and ASR applications	no	-
AlfaNum Speech Databases for TTS	AlfaNum TTS	alfanum.ftn.uns.ac.rs	corpus	18	hour	monolingual	yes	no (3)	yes	For TTS and ASR applications	no	-
Digital Archive of the Institute for Balkan Studies	DABI	www.balkaninstitut.com	corpus	2000	hour	monolingual	yes	no (3)	yes	For TTS and ASR applications	no	-
Corpus of Serbian Language	CSL	www.serbian-corpus.edu.rs/ns/eindex.htm	corpus	11000000	word	monolingual	yes	no (3)	no	-	no	-
Organizing digitized material	InfoBeaver	http://www.korpus.matf.bg.ac.rs/InfoBeaver	tool	-	-	-	no	1	yes	An application for collecting and presenting multimedia informations.	yes	Published in 1st batch
Second D2.3												
Anthology of Serbian Literature	ASK	www.ask.rs	corpus	130	file	monolingual	yes	no (3)	yes	Collection of Serbian literary texts	no	-
Media Archive Ebart	EbartArchive	http://www.arhiv.rs/	corpus	3.3 million	article	monolingual	yes	2	yes	Large corpus of newspaper texts	yes	-
English-Serbian Corpus of Abstracts of Scientific Projects	SrpEngSciKor	-	corpus	350 000	word	bilingual	yes	no (3)	yes	Bilingual corpus of scientific texts from various domains	yes	-
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	yes	no (3)	yes	Spelling check for Serbian	yes	-
Morphosyntactically tagged Serbian version of Jules Verne's novel "Around the world in 80 days"	Verne80daysMSD	http://www.korpus.matf.bg.ac.rs/Verne80daysMSD	text	58676	word	monolingual	no	2	yes	Morphosyntactically tagged, lemmatized and disambiguated text conforming to MULTEXT-East	yes	Published in 2nd batch

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
Aligned Collection Search Tool	Biblisha	http://hlt.rgf.bg.ac.rs/Biblisha	tool	-	-	-	no	2	yes	A web application for search of digital libraries of articles from bilingual e-journals	yes	Published in 2nd batch
Serbian module for Nooj	SrpNooj	http://www.nooj4nlp.net/pages/resources.html	corpus, lexical / conceptual resource	-	-	-	no	2	yes	A set of resources necessary for using Nooj for Serbian (including Serbian MD)	yes	Published in 2nd batch
Third D2.3												
Multimedia Ebart Archive	EbartArchive	http://www.arhiv.rs/	speech corpus	-	article	monolingual	yes	3	yes	Large corpus of multimedia broadcast material transcribed to text	yes	Negotiate license agreement
Terminological Database for Geology	GeolISSTerm	http://www.rgf.bg.ac.rs/	lexical / conceptual resource	3500	concepts	bilingual	yes	no (3)	yes	Bilingual terminological database	yes	Negotiate license agreement
Serbian-English Aligned Literary Corpus		http://www.ff.uns.ac.rs/	corpus	-	word	bilingual	yes	3	yes	Aligned corpus of Serbian texts translated to English	no	-
Named entities evaluation corpus for Serbian	SrpNE-evaluation	http://korpus.matf.bg.ac.rs	corpus	150000	word	monolingual	no	3	yes	Corpus automatically tagged and manually checked with various NE	yes	Clean-up
Named entities module for Serbian	SrpNE-module	http://korpus.matf.bg.ac.rs	tool	-	-	monolingual	no	no (3)	yes	A FST module for shallow parsing of NE	yes	Negotiate license agreement
Language model for Serbian	-	http://www.arhiv.rs/	lexical resource	-	-	monolingual	yes	no (3)	yes	Serbian language module built upon large corpus data	yes	Negotiate license agreement
Web applications (NE extraction from web pages)	-	http://korpus.matf.bg.ac.rs	tool	-	-	multilingual	no	no (3)	yes	A web tool for extraction of proper names from Wikipedia	yes	Negotiate license agreement
Emotion classification of Serbian Texts	-	http://korpus.matf.bg.ac.rs	tool	-	-	monolingual	no	no (3)	yes	A web tool for sentiment analysis of Serbian texts	yes	Under development
Semantically tagged Corpus of Contemporary Serbian (preliminary version)	-	http://korpus.matf.bg.ac.rs	corpus	-	word	monolingual	no	3	yes	Corpus semantically tagged on the basis of semantic attributes in e-dictionaries	yes	Under development

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	LinguisticType	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
A web tool for aligned text search	-	http://korpus.matf.bg.ac.rs	tool	-	-	multilingual	no	no (3)	yes	A web tool for effective search of aligned and annotated texts	yes	Under development

7.6. Slovak

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	Linguality Type	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
First D2.3												
Slovak National Corpus	prim	LSIL	corpus	7,19E+08	token	monolingual	no	1	yes	A representative corpus of contemporary Slovak language written texts published since 1955 (1953 being the time of most recent substantial Slovak language orthography reform).	yes (as a query interface)	Published in 1st batch.
Corpus of Spoken Slovak	hovor	LSIL	corpus	2600000	token	monolingual	no	1	yes	Provides material for research and description of the real form of contemporary standard spoken Slovak.	yes	Published in 1st batch.
Slovak Morphological Lexicon	ma	LSIL	lexicalConceptualResource	77000	lemma	monolingual	no	1	yes	Foundation of any reasonably complex NLP analysis	yes	Published in 1st batch.
Slovak Treebank	Slovak Treebank	LSIL	lexicalConceptualResource	50000	sentence	monolingual	no	2	yes	Slovak language treebank of manually syntactically annotated sentences.	yes	Published in 1st batch.
Slovak WordNet	wn	LSIL	lexicalConceptualResource	12500	synset	multilingual	no	3	yes	Basic ontology database.	yes	Tool development
Slovak-English Parallel Corpus	sk-en	LSIL	corpus	1500000	sentence	bilingual	no	1	yes	Big parallel bilingual corpus.		Published in 1st batch.
Slovak-Czech Parallel Corpus	sk-es	LSIL	corpus	700000	sentence	bilingual	no	1	yes	Big parallel bilingual corpus.	yes (as a query interface)	Published in 1st batch.
Slovak Web Corpus	Sk-web	LSIL	corpus	1E+09	token	monolingual	yes (partly)	2	yes	Web corpus contains texts downloaded from the .sk domain. The texts are automatically lemmatized and morphologically tagged.	yes (as a query interface)	Published in 2nd batch.

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	Linguality Type	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
Slovak Legal Texts Corpus	legal	LSIL	corpus	1,46E+08	token	monolingual	yes (partly)	2	yes	Corpus of legal texts contains legal regulations and other available legal documents (laws, decrees, announcements, directives, protocols, etc.) The corpus has been prepared in collaboration with the Ministry of Justice of the Slovak Republic.	yes (as a query interface)	Published in 2nd batch.
Slovak-Russian Parallel Corpus	sk-ru	LSIL	corpus	100000	sentence	bilingual	yes (partly)	no	no	Big parallel bilingual corpus.	yes (as a query interface)	Tool development,
Slovak-French Parallel Corpus	sk-fr	LSIL	corpus	21000	sentence	bilingual	no	no	no	A parallel bilingual corpus of an important language.	yes (as a query interface)	Tool development
Second D2.3												
Balanced Slovak Corpus	VYV	LSIL	corpus	2,47E+08	token	monolingual	no	2	yes	A balanced corpus with respect to text type. It contains 1/3 fiction, 1/3 informational text, 1/3 professional text (including popular science).	yes (as a query interface)	Published in 2nd batch.
Dictionary of Slovak Noun Collocations	Dictionary of Slovak Noun Collocations	University of St. Cyril and Methodius, Trnava	Lexical / Conceptual Resource	250	entry	monolingual	yes (partly)	no	no	A manually lemmatized and morphosyntactically annotated corpus. It is used as a basis for NLP tools training (primarily POS tagger and lemmatizer).	no	Negotiate license agreement.
Manually Annotated Slovak Corpus	MAK	LSIL	corpus	1200000	token	monolingual	no	2	yes	Necessary for training of statistical tools	yes (query interface)	Published in 2nd batch.
Slovak Terminology Database	STD	LSIL	Lexical / Conceptual Resource	4500	entry	monolingual	no	3	yes	High quality database of terms from various domains	yes	Add term annotations, proofread existing terms.
Language model prim-5.0-sane	Language model prim-5.0-sane	LSIL	Technology/tool	7,33E+08	token	monolingual	no	2	yes	A language model from the Slovak National Corpus.	yes	Published in 2nd batch.
Language model prim-5.0-inf	Language model prim-5.0-inf	LSIL	Technology/tool	5,15E+08	token	monolingual	no	2	yes	A language model of journalistic style.	yes	Published in 2nd batch.

resourceTitle	resourceName	resourceLocation	resourceType	size	sizeUnit	Linguality Type	Outside the consortium	Published in META-SHARE	Important to be published in META-SHARE	Describe why the resource is of further interest	Willing to publish resources in the META-SHARE	Describe the transformation work need to be done to make resources publishable
Language model prim-5.0-vyv	Language model prim-5.0-vyv	LSIL	Technology/tool	2,47E+08	token	monolingual	no	2	yes	A language model of balanced language built on the balanced Slovak corpus.	yes	Published in 2nd batch.
Third D2.3												
Database of root morphemes	Database of root morphemes	Prešov University	Lexical / Conceptual Resource	67000	root morpheme	monolingual	yes	no	no	Provides alternative approach to morphology analysis	-	Negotiate license agreement.
Dictionary of Slovak Adjective Collocations	Dictionary of Slovak Adjective Collocations	University of St. Cyril and Methodius, Trnava	Lexical / Conceptual Resource	70	entry	monolingual	yes	no	no	Overview of the combinatorial behaviour of words – contains collocation profiles of the most frequent Slovak adjectives	-	Negotiate license agreement.
Dictionary of German-Slovak Collocations	Dictionary of German-Slovak Collocations	University of St. Cyril and Methodius, Trnava	Lexical / Conceptual Resource	40	entry	bilingual	yes	no	no	Overview of the combinatorial behaviour of words in bilingual comparison	-	Negotiate license agreement.
Multimodal multilingual dictionary of gestures: DiGest	DiGest	Institute of Informatics, Slovak Academy of Sciences	Lexical / Conceptual Resource	324	entry	multilingual	yes	no	no	Contains database of extra-verbal expressions	-	Negotiate license agreement.