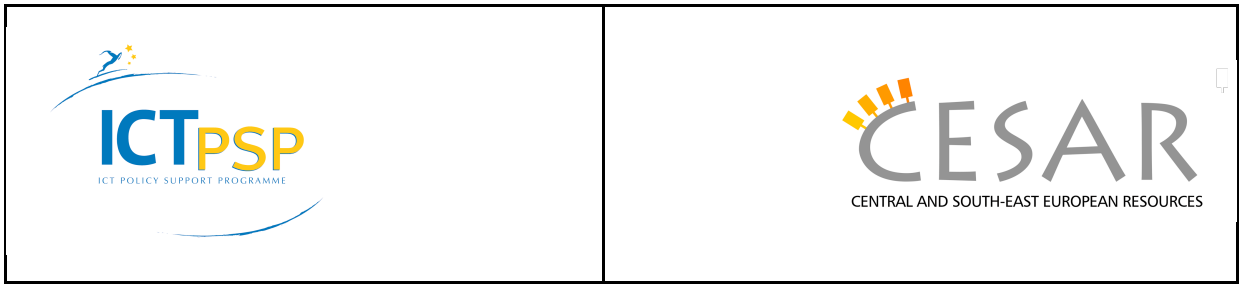

CESAR

Central and South-East European Resources

Project no. 271022

Deliverable D3.3.-B

Third batch of language resources: actions on resources

Version No. 1.1

07/02/2013

Document Information

Deliverable number:	D3.3.-B
Deliverable title:	Third batch of language resources: actions on resources
Due date of deliverable:	31/01/2013
Actual submission date of deliverable:	08/02/2013
Main Author(s):	György Szaszák (BME-TMIT)
Participants:	Mátyás Bartalis, András Balog (BME-TMIT) Željko Agić, Božo Bekavac, Nikola Ljubešić, Sanda Martinčić-Ipšić, Ida Raffaelli, add Jan Šnajder, Krešo Šojat, Vanja Štefanec, Marko Tadić (FFZG) Cvetana Krstev, Duško Vitas, Miloš Utvić, Ranka Stanković, Ivan Obradović, Miljana Milanović, Anđelka Zečević, Mirko Spasić (UBG) Svetla Koeva, Tsvetana Dimitrova, Ivelina Stoyanova (IBL) Tibor Pintér (HASRIL) Mladen Stanojević, Mirko Spasić, Natalija Kovačević, Uroš Milošević, Nikola Dragović, Jelena Jovanović (IPUP) Anna Kibort, Łukasz Kobyliński, Mateusz Kopeć, Katarzyna Krasnowska, Michał Lenart, Marek Maziarz, Marcin Miłkowski, Bartłomiej Nitoń, Agnieszka Patejuk, Aleksander Pohl, Piotr Przybyła, Agata Savary, Filip Skwarski, Jan Szejko, Jakub Waszczuk, Marcin Woliński, Alina Wróblewska, Bartosz Zaborowski, Marcin Zając (IPIPAN) Piotr Pęzik, Łukasz Dróżdż, Maciej Buczek (ULodz) Radovan Garabík, Adriána Žáková (LSIL)
Internal reviewer:	HASRIL
Workpackage:	WP3
Workpackage title:	Enhancing language resources

Workpackage leader:	IPIPAN
Dissemination Level:	PP
Version:	1.0
Keywords:	Upload, interoperability, standardization, harmonization, upgrade, extension, linking

History of Versions

Version	**Date**	**Status**	**Name of the Author (Partner)**	**Contributions**	**Description/ Approval Level**
1.1	08/02/2012	final	Tamás Váradi (HASRIL)		proofreading
1.0	07/02/2012	final	Tibor Pintér (HASRIL)		proofreading
0.9	06/02/2013	final	György Szaszák (BME-TIMIT)		editing, cross-checking
0.8	05/02/2013	final	Tibor Pintér (HASRIL)		editing
0.7	19/01/2013	draft	György Szaszák (BME-TMIT), Maciej Ogrodniczuk (IPIPAN)	From all Partners	

Executive summary

This deliverable entitled D3.3.-B. is a supplement for the publicly available deliverable D3.3., treating third upload batch of language resources and tools. In this deliverable emphasis is put on the description of the work and activities related to each and every resource included in the third upload batch. Work for resources or tools uploaded in the first or second batch, but extended or updated for the third batch is also documented.

Table of Contents

Table of Contents

Abbreviations

Abbreviation	Term/definition
LR	Language Resource
LRT	Language Resources and Tools (either language data or tools)
Partners	Partners of CESAR
KPI	Key Performance Indicators
IPR	Intellectual Property Rights

Table 1. Abbreviations

0. Scope

0.1. Actions – upgrading resources

Tasks 3.1, 3.2. 3.3 in the DoW are linked to upgrading resources to agreed standards, by focusing on reaching META-SHARE compliance, which in some cases may need additional actions, depending on the tool/resource.

The foreseen activities by the DoW were:

- upgrade for interoperability (changing annotation format, type, tagset),
- technology-related upgrade (wrapping, refactoring, etc.),
- metadata-related work (creation, enhancement, conversion, standardization),
- harmonization of documentation (conversion to open formats, reformatting, linking),
- preparation for maintenance and deployment (debugging, cleaning, building test environments, preparing code repositories), programming tasks (bug-fixing and standardizing API calls).

By the selection of the resources to be upgraded, the following principles were kept in mind:

- the resources are state-of-the-art representatives of their type for a certain language,
- if more than one valuable representative of certain tool type for a language is available (e.g. two morphosyntactic analysers with equally popular tagsets or formal grammars used for different purposes), all of them are included in the selection,
- current status of resources present superior quality at least on regional level without the need of excessive further development,
- licensing issues allow to process and make available the resources and resource-related materials as free as and as open as possible, depending also on the fact whether the consortium succeeds in reaching an agreement with respective copyright holders.

More details on the selection of resources have already been given in deliverables D2.3a-c.

0.2. Actions – extending and linking resources

Tasks 3.2. and 3.3., which have higher importance in this batch are linked to extension, linking of resources. For resources released in the first batch, the – not exclusive – emphasis was put on upgrading, interoperability and META-SHARE compliance, whilst for the second and third batches, extension and linking of the selected resources and tools had already come

to a maturity level which allows for their upload onto the META-SHARE nodes. This might affect the extension of batch 1 or 2 resources in the present batch too.

Existing resources may need to be extended or linked across different sources to improve their coverage and increase their suitability for both research and development work. Task 3.2 takes into account the specific goals of the project, identified gaps in the respective language community, and most relevant application domains.
Selection of resources to be extended/linked based on those made available within task 3.1 to further enhance a smaller, but well-defined set of resources. Following rationale was applied:

- the extension of resources should provide considerable value to the community, at least on regional level,
- the emphasis is on providing building blocks to the existing tools (e.g. extended grammars to existing shallow parsers) rather than major restructuring,
- additional resources are integrated with existing ones only if they significantly improve the quality of resulting resources,
- if more than one representative of certain tool type for a language has been selected in task 3.1, they are very likely to be interlinked to benefit from strong points of both solutions (unless their usage patterns do not encourage such action),
- if less-developed, but still very popular (at least within one language community) tools can benefit from the enhancement basing on their well-developed equivalent (provided that no extensive work will be necessary and that the latter tool cannot be used as a building block in further applications of the former tool), their enhancement was also considered,
- experience of other consortium members (or, where applicable, other consortia) is extensively used in the process of further extending national resources to provide strong foundation for cross-linguality,
- tools offering language-neutrality or cross-linguality are preferred.

More details on the selection of resources for linking and extension have already been given in deliverables related to WP2.

0.3. Actions – Aligning resources across languages

Cross-lingual alignment of resources, as the most demanding task, was be applied only to a small number of resources. The following rationale was applied:

- no more than a tool of a certain type for each language tuple is used in the process,
- whenever applicable, the largest set of languages is selected (preferably with English as a hub language; the languages going beyond natural consortium scope of interest are not excluded),
- language-independence is targeted to a great extent,

- the quality of a result is of immense concern (not the quantity of the integrated tools), which will be assessed according to standard evaluation measures used for LRs.

0.4. Management of the 3rd batch

The 3rd upload batch and related activities were grouped in CESAR as follows:

- set up META-SHARE node version 3.1, provide metadata schemes version 3.0. and IPR requirements for CESAR Partners – in close collaboration with META-NET and partner projects – and pilot service – this action is presented in D4.4.
- metadata description, resource documentation (documentation of the delivery from resource point of view) – presented and listed in D3.3.
- actions on resources as required in WP3 (upgrade, extension, standardization, harmonization, etc.) – presented here in D3.3.-B.

This document is an extension for D3.3. deliverable and its paragraph numbering follows that of D3.3. This means that related activities to a given LRT described under a given section in D3.3. will be presented here in the same section.

The following sections provide a bullet-point style overview of the activity for each LRT in batch 3. If some explanation was found necessary, it is also provided, accompanied by the arguments which justify eventual non-foreseen or otherwise foreseen activities related to the LRT.

Partners of CESAR have not only fulfilled, but even overperformed their obligations, as some resources scheduled for later batches have already been included in batch 1 and now in batch 2. This made possible to add some extra LRTs to batch 3 and carry out further actions on the already released LRTs – which may contribute to considerably enlarge their usability (be on a higher level of technology, wider range of usability, etc.).

1. HASRIL resources

1.23. Hungarian National Corpus

The national corpus of Hungarian language, which is derived into five subcorpora by regional language variants, and into five subcorpora by text genres also. The subcorpus to be studied can be chosen by any combination of these. That makes the HNC an appropriate tool to study the differences not just between text genres but between language variants. HNC wishes to be a representative general-aim corpus of present-day standard Hungarian.
The following work was carried out within the CESAR project:

- updated tools for analysis and annotation
- higher quality and finer level of analysis and annotation (detailed morphosyntactic analysis and disambiguation with updated processing toolchain, NP chunking, Named Entity recognition, distributional analysis, built in post-processing (multilevel frequency lists, subsequent searches on previous results))
- new technology for development
- extension to 1 GW
- new samples of language use and include further variants (transcribed spoken language data in particular)
- cleaned IPR issues of the involved texts
- extension of metadata

1.24. HUCOMTECH multimodal database

The HuComTech multimodal corpus consists of about 50 hours of video and audio recordings of 111 formal dialogues (simulated job interviews) and 111 informal but guided dialogues. The language of the recordings is Hungarian. The participants were university students aged 19-27, female 54 and male 67. The corpus was annotated for video (facial expressions, instances of eyebrows, gaze, headshift. handshape, touchmotion and posture) and audio (emotions, discourse, prosody and textual transcriptions). Its unique features in a wider comparison include its special attention to pragmatics focusing on a comparative study of the unimodal vs. multimodal features of communication (as compared to multimodality alone) as well as the study of the syntax and prosody of spoken language with respect to the above wide range of multimodal characteristics. The data can be queried in ELAN and in our web-based SQL database.
The following work was carried out within the CESAR project:

- Conversion of the annotations to ELAN (.eaf) format
- Continuation of the multimodal and unimodal pragmatic annotation
- Continuation of the syntactic annotation

1.25. BEA Hungarian spontaneous speech database

The aim of developing a phonetically-based multi-purpose database of Hungarian spontaneous speech, dubbed BEA (BEszélt nyelvi Adatbázis 'spoken language database'), is to accumulate a large amount of recorded spontaneous speech produced by numerous present-day Budapest speakers, providing ample material for various types of research and practical applications.

At the time of writing, the total recorded material of BEA is 250 hours long, meaning approximately 3,400,000 running words. The database primarily contains spontaneous speech materials, but for the sake of comparisons, it also includes sentence repetitions and read texts.

The database offers material for research in a number of areas within linguistics. The study of acoustic-phonetic consequences of the production of speech sounds, coarticulation effects, and suprasegmental features was hampered for many years by the methodological difficulty that no spontaneous speech material of an adequate quality was available. In addition to phonetic research in the strict sense, now it becomes possible to carry on conversation analysis, pragmatic research, speech technology, the study of speech accommodation, that of the spontaneous speech of elderly speakers, or that of disfluency phenomena.

The following work was carried out within the CESAR project:

- anonymization of sound files and transcription
- the sound file transcriptions of BEA materials are done at several levels
- primary transcription in orthography but without punctuation. Transcribers use Microsoft Office Word (.doc format)
- annotation: This form of transcription is a kind of visual display of spoken texts and some further pieces of information related to them in a way that the written text and the actual recording can be displayed/listened to simultaneously. This is made possible by software Transcriber.

1.26. Hungarian kindergarten language corpus

The Hungarian Kindergarten Language Corpus (HUKILC) has been compiled predominantly for child language variation studies. It contains 62 interviews with 4,5-5,5 year-old kindergarten children from Budapest. The interviews are at least 20 minutes long. HUKILC contains cca. 192.000 words. The children are divided into 4 groups concerning socio-economic status (SES) and sex. There is a higher SES group with males (hm), and with females (hf), and a lower SES group with males (lm) and females (lf), respectively. The corpus is also a useful source for other fields of child language research (eg. phonetics, or developmental morphology).

The following work was carried out within the CESAR project:

- sound files transcription using CHAT format of CHILDES
- anonymization of transcripts
- morphological analyzer (Humor) and disambiguator (PurePos) adapted for child language data (still in progress)

1.27. ht-online (lexical resource of the Hungarian outside Hungary)

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 spacial lexical resource in the Hungarian language tools based on the Hungarian morphology.
The following work was carried out within the CESAR project:

- Update of the SQL-database behind
- Update on the query structure (more enhanced)
- Minor bug-fixing in the engine

1.28. Hungarian concise dictionary (with sample sentences)

A unique dictionary of Hungarian language of 16 000 headwords (entries) followed by frequency data. Each entry describes the most common forms (given by pragmatic reasons) of the headword. The entries are divided into meanings which counts 33 000 carefully selected and stylistically labeled meanings. The dictionary contains sentences brought from real language use and 3000 phrasems.
The following work was carried out within the CESAR project:

- XML-conversion of the dictionary
- documentation

1.29. Hungarian historical corpus

Hungarian historical corpus (further as HHC) is a collection of texts written between 1772 and 1997 in different genres, containing ca. 27 million tokens. During the compilation of HHC, text samples were selected by professionals (literary historians, historians, mathematicians etc.) from printed works. A relative majority (40%) of the texts are dated from the second half of the 20th century. The corpus is the product of the Department of Lexicography and Lexicology at RIL HAS, made between 1986 and 1997, maintained continuously since then. As an innovation, genre labeling was unified. Thus, genres and text types in HHC and HNC are marked similarly, this makes possible to search data of these corpora by using the same query structure.

The following work was carried out within the CESAR project:

- shared query structure (HHC and HNC)
- genre labelling

1.30. N-grams from Hungarian National Corpus

N-GRAM lists of the current version of the Hungarian National Corpus (before the upgrade) containing 229 million units (the HNC contains 187 million tokens, but for the N-Grams we used the corpus with punctuation). The N-grams were provided as:

- N-Grams made on stems and tokens
- N-Grams: 2-Grams, 3-Grams, 4-Grams, 5-Grams

1.31. Corpus of Hungarian school metalanguage

Corpus of Hungarian school metalanguage (CHSM) is a corpus of semi-structured research interviews collected during a fieldwork carried out by Tamás Péter Szabó in 2009 as an individual project. This corpus contains 74 interviews recorded with students (aged 6–11, 13–15 and 17–19) and teachers of Hungarian grammar and literature. Topics cover narratives and ideologies on repair strategies, evaluation of other speakers' practice (dialects, slang, curse, impediment etc.), narratives on grammar courses and others. CHSM, containing 346,500 words, annotated manually is suitable for Conversation Analysis and Discourse Analysis investigation of ideology constructions emerging in Hungarian metadiscourses. CHSM contains the transcription of recorded speech, voice recordings are planned to add later. (Because of ethical issues, in several cases, the whole text of the interview is not available.)

The following work was carried out within the CESAR project:

- conversion of XML files following TEI 5 recommendations
- morphological analysis using HUMOR
- new query structure supporting public use

2. BME-TMIT resources

2.14 Hungarian Book (Egri csillagok/Eclipse of the Crescent Moon by Géza Gárdonyi) Reading Speech and Aligned Text Selection Database

For the 3rd batch the following work has been carried out:

- Metadata updated

2.15 Hungarian Poem (János vitéz/John the Valiant by Sándor Petőfi) Reading Speech and Aligned Text Selection Database

For the 3rd batch the following work has been carried out:

- Metadata updated

2.16 Hungarian Parliamentary Speech and Aligned Text Selection Database

The Hungarian Parliamentary Speech and Aligned Text Selection Database has already been included in the 2nd batch.
For the 3rd batch the following work has been carried out:

- All the speeches available on the Hungarian Parliament website have been downloaded and processed. The size of the database increased by more than 10 times.
- The speech recognition and sample selection method has been improved to produce more accurate samples in the selection corpus.

2.17 Named entity lexical database

The Named entity lexical database contains spoken Hungarian given names, names of towns and townships, and pronounced forms of publicly owned lands (street, square, etc.) needed for the correct postal addresses.
Actions carried out within CESAR involved:

- Checking and optimizing the waveforms of the word items
- Checking the database elements and the unified structure
- Filling and finalizing METADATA schemes and files

2.18 Hungarian formant database

The goal of the precisely annotated and segmented formant database is to give the formant data in Hungarian vowels at three points inside the vowel. The formants are determined in spoken words and the formant values are stored in an Excel spreadsheet file. The database can be used for speech research, development and in education showing examples for formant values and trajectories according to the sound environment.

Actions carried out within CESAR involved:

- Checking and optimizing the waveforms of the word items
- Automatic marking of sound boundaries and sound symbols on the waveform
- Manual checking of the marked sound boundaries and the given sound symbols
- Checking the database elements and the unified structure
- Automatic measuring of F1,F2,F3 and F4 formnat values on three measuring points within the vowels
- Manual checking of the formant values and correcting them
- Filling and finalizing METADATA schemes and files

2.19 Graphical query interface for Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection

This tool supports the visual study and correction of the formant values derived from the ParallelSpeech-hu database. It gives the picture of the spectrum of the given sentence, the sound boundaries and sound symbols, the place of measured formant values in 5 measuring points inside the vowel. The tool makes possible of searching according to sound environment, filtering the database items (eg. male /female) and interactive correction of the numerical formant values (in case of wrong formant measurement). The tool gives only visual information, sound not.

Actions carried out within CESAR involved:

- Checking and optimizing the waveforms of the word items
- Automatic marking of sound boundaries and sound symbols on the waveform
- Manual checking of the marked sound boundaries and the given sound symbols
- Checking the database elements and the unified structure
- Automatic measuring of F1,F2 and F3 formant values on five measuring points within the vowels and organizing them into a database
- Manual checking of the formant values and correcting them
- Automatic making the pictures about the spectral shape of each sentences
- Developing the graphical interface to view the spectral pictures and formant values together in the same screen
- Development of searching and filtering algorithms for interactive use
- Filling and finalizing METADATA schemes and files

2.20 Medical Speech Database

Hungarian Medical Speech Database is a newly created and continuously enhanced speech database which contains pathological speech uttered by speakers suffering from various

speech disorders. The development of the database is carried out within CESAR involving mainly:

- Database creation, transcription (orthographic)
- Providing classification for underlying medical state of the speaker
- Partial segmentation (phone level, semi automatic, hand-checked)
- Structurization, format supporting interoperability
- Documentation preparation
- Clearance of license terms
- Metadata description

2.21. Automatic Prosodic Segmenter

Automatic Prosodic Segmenter is a tool designed for prosodic segmentation of speech utterances down to the phonological phrase level if possible. The system is designed to be language independent, however, it highly relies on fixed stress, and hence, the range of other languages for which the system is adaptable can be restricted. Hungarian read speech phonological phrase models are provided, but the system is retrainable using the HTK toolkit for other languages provided that data is available. The actions carried out within CESAR included:

- Bug-fix and enhancement
- Documentation prepareation
- Metadata annotation
- IPR clearence

2.22. Hungarian Phonetic Transcriber

Hungarian Phonetic Transcriber is a rule based application designed to perform automatic phonetic transcription. As written and spoken language are close in standard formal Hungarian, the tool is highly reliable, however, when intended to used for the transcription of proper nouns, a dictionary holding exceptions and special pronunciations is required. This is not included but can be compiled from other resources available in CESAR.

Actions carried out within CESAR:

- Documentation improvement
- Bug-fix and upgrade
- Metadata annotation
- IPR clearence

2.23 Hungarian MALACH Speech Database

The Hungarian MALACH Speech Database is created as a part of the MALACH (Multilingual Access to Large Spoken Archives) database. It contains 116,000 hours of

digitized interviews in 32 languages from 52,000 survivors, liberators, rescuers and witnesses of the Nazi Holocaust.

The presented hungarian database includes 24 whole interviews and parts of 104 interviews with Hungarian speakers.

The recordings are segmented between speech pauses for speech recognition software development purposes. The interview parts should be used to train the recognizer, the complete interviews are chosen for testing.

Manual transcriptions for all the audio files are also included in the database.

For the 3rd batch the following work has been carried out:

- database cleaning
- metadata creation

2.24 Hungarian Broadcast Conversation Database from the Catholic Radio of Eger (EKR)

The recordings are mostly complete radio programmes, and contain semi-spontaneous conversations.

For every sound file there is a transcription text, which is created with the Transcriber tool manually. The document of the transcription rules is also attached to the database.

For the 3rd batch the following work has been carried out:

- database cleaning
- metadata creation

2.25 Accent marker database for Hungarian written sentences

The goal of this database is that it gives information on text level, whether the word in the sentence is accented during pronunciation or not. These accent markers may help the testing of accent modeling algorithms and other language technology research.

Actions carried out within CESAR involved:

- Automatic definition of accent markers on words using the accent module of a Hungarian TTS
- Manual checking of the predefined accent markers in each sentences
- Audio checking the final accent marker distribution by a data-driven TTS
- Filling and finalizing METADATA schemes and files

3. FFZG resources

3.13. Croatian Language Web Services (hrWS)

The following work has been carried out for the resource within CESAR:

- the rule-based sentence splitter was produced in the form of FST
- the rule-based tokeniser was produced in the form of php-scripts
- the PoS/MSD-tagger and lemmatiser was produced in the form of FST
- the NERC module was produced using StanfordNERC engine
- the first Croatian dependency parser was produced in the form of a hybrid parser
- a wrapper for all modules was designed and produced
- the web services web front end has been produced (http://lt.ffzg.hr)
- compilation of metadata and description was produced.

The Croatian Language Web Services (hrWS) is a set of language processing web services oriented towards the processing of Croatian language by evoking different modules. The modules supported with this version of hrWS are:

- sentence splitting
- tokenization
- PoS/MSD-tagging
- NERC
- dependency parsing.

The hrWS use standard REST protocol for input and output of data to process and can be accessed at http://lt.ffzg.hr.

3.14. Croatian Translations of Acquis Communautaire (hrAcquis)

The following work has been carried out for the resource within CESAR:

- the conversion from MS-DOC format into XML following the JRC-Acquis DTD was produced
- the whole corpus was sentence splitted
- the whole corpus was sentence aligned using HunAlign
- TMX formatting produced
- XML and TMX available for download
- compilation of metadata and corpus description was produced

The Croatian Translations of Acquis Communautaire is a bilingual English-Croatian parallel corpus that is following the specification set out by the JRC-Acquis Parallel Corpus. This corpus is composed of 16,638 documents translated within the Translation Service of the

Ministry of Foreign Affairs and European Integrations of the Republic of Croatia until 2012-08-28. The conversion from .doc to XML was following the JRC-Acquis DTD. The English part of the Corpus was extracted from the original JRC-Acquis Corpus using CELEX codes to filter out the needed documents. The English-Croatian sentence alignment was processed with HunAlign alignment tool. The corpus is distributed under the CC-BY-SA licence.

3.15. Croatian National Cropus v3.0 (HNK v3.0)

The following work has been carried out for the resource within CESAR:

- crawling and conversion of Vjesnik (2004-2010)
- conversion of Vijenac (2002-2005)
- adding of new texts (Vjesnik, Vijenac, Narodne novine) to the existing HNK v2.5
- migration of the corpus to Bonito2 (or NoSketchEngine) corpus storage and search engine (http://filip.ffzg.hr/bonito2/run.cgi/first_form)
- compilation of metadata and corpus description.

The Croatian National Corpus v3.0 (HNK v3.0) is a representative corpus of contemporary Croatian standard language written texts published since 1990. The corpus is automatically lemmatised and MSD tagged. The documents are annotated with their genre, type and other information. The whole corpus is composed of faction, fiction and mixed texts. This is a pseudocorpus, only the query interface using Bonito2 web interface is available, while the original texts cannot be distributed for copyright reasons. Bonito2 web interface gives opportunities to issue complex queries due to elaborated query language resulting not only in concordances, but also in word-lists, collocations and other types of distributional data etc. of tokens, lemmas and/or MSDs. This version of HNK features Bonito2 web interface and additional texts.

3.16. Corpus of Narodne novine (NNCorp)

The following work has been carried out for the resource within CESAR:

- the whole corpus was recrawled until 2012
- conversion from different HTML formats to a unique XML
- the corpus was automatically PoS/MSD-tagged and lemmatised
- the corpus was stored for access with Bonito2 (http://filip.ffzg.hr/bonito2/run.cgi/first_form?corpname=nn)
- compilation of metadata and corpus description

The Corpus of Narodne novine v2.0 (NNCorp v2.0) is a specialised corpus of contemporary Croatian standard language written texts published since 1990 in the official journal of the Republic of Croatia "Narodne novine". The corpus is automatically lemmatised and MSD tagged. This is a pseudocorpus, only the query interface using Bonito2 web client is available, while the original texts cannot be distributed. Bonito2 web interface gives opportunity to

issue complex queries, thanks to elaborated query language, resulting not only in concordances, but also in word-lists, collocations and other types of distributional data etc. of tokens, lemmas and/or MSDs. The NNCorp v1.0 was entirely included in HNK v2.5 and this v2.0 is entirely included in HNK v3.0.

3.17. Croatian n-grams (hrNgrams)

The following work has been carried out for the resource within CESAR:

- computing the n-grams (1-, 2- and 3-grams) of the Croatian National Corpus v2.5
- compilation of metadata and resource description

This resource contains sets of n-grams of different sizes (from 1 to 3) computed from the Croatian National Corpus v2.5. N-grams were computed both from lowercased text and text in original character case. For every size of n above one (i.e. for bigrams and trigrams), n-grams were computed in two ways: taking to account only those appearing within sentence and across sentence boundaries. Regarding the tokenization of the corpus, token is considered to be a continuous sequence of non-whitespace characters. Punctuation markings are treated as separate tokens. Complex punctuations are tokenized as a sequence of simple punctuations. Resource consists of 10 textual files, each computed with different combination of paramaters (i.e. n-gram length, character case, sentence boundaries). Each line in the file represents one unique n-gram and its absolute frequency in the corpus, separated by a tabulator. N-grams are ordered according to their frequency, starting from highest to lowest. The n-grams lists were produced using methodology and tools developed by the CESAR Polish partner IPIPAN.

3.18. Croatian Morphological Lexicon v5.0 (HML5)

The following work has been carried out for the tool within CESAR:

- the previous version of Croatian Morphological Lexicon (v4.6) was matched with the Croatian National Corpus (v2.5) and tokens unknown to the Lexicon were collected
- for collected lemmas inflectional patterns were attributed by hand
- for collected lemmas all wordforms were generated
- all newly generated wordforms were added to the existing v4.6 thus producing v5.0
- compilation of metadata and description

The Croatian Morphological Lexicon is an inflectional lexicon generated automatically by Croatian Inflectional Generator from ca 125,000 lemmas yielding over 5,000,000 word forms. It has been a result of the group lead by Marko Tadić on the basis of theoretical background published in 1992 (see Tadić 1994). The initial set of lemmas was collected from several existing Croatian mono- and bi-lingual dictionaries, while additional entries were collected via corpus or by means of automatic enlargement of the initial list of lemmas (see Bekavac, Šojat 2005, and Oliver, Tadić 2004 below). The automatically generated output was corrected for known systemic errors, encoded in utf-8 and stored in MulTextEast Lexica format: lemma[TAB]word-form[TAB]MSD. The MSD-tagset is conformant with the MulTextEast

v4.0 reccomendations for Croatian language. However, some additions exist: in surnames gender is left unspecified (-), additional subclassification of adverbials has been introduced etc. At the moment the Croatian Morphological Lexicon is a distributed under CC-BY-NC-SA license.

3.19. Orwell 1984 Croatian (hr1984)

The following work has been carried out for the tool within CESAR:

- conversion of MS-DOC into XML format
- automatic PoS/MSD-tagging and lemmatisation produced
- PoS/MSD-taggin and lemmatisation manually corrected
- compilation of metadata and description

The Orwell 1984 Croatian is a Croatian contribution to the MULTEXT-East resources, a multilingual dataset for language engineering research and development. This dataset contains linguistically annotated translations of Orwell's novel 1984 in Bulgarian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Romanian, Serbian, Slovak, Slovene. This corpus adds the Croatian version to the set. The texts in this corpus are lemmatized, MSD-tagged and manually corrected following MTE v4.0 specifications.

3.20. Croatian Wordnet (CroWN)

The following work has been carried out for the tool within CESAR:

- translation of BCS1, 2 and 3
- manual proofing and correction of translations
- installing the Hydra wordnet management system
- additional development of Hydra
- compilation of metadata and description

The Croatian wordnet (CroWN) is a semantic network of Croatian lexis. CroWN was built on the basis of Princeton WordNet v2.0 and synchronized at the end with the Princeton WordNet v3.0. Thus, CroWN is completely compatible with PWN 3.0 and, consequently, with all other wordnets mapped to it. It comprises 31,300 literals in 10,040 synsets. Out of all literals, 16,757 (53.54%) of them are nouns, 13,680 (43.7%) verbs, 857 (2.73%) adjectives and 6 (0.02%) of them are adverbs. As such, CroWN covers 98.87% of synsets from BCS 1, 2 and 3. This resource contains three export files from Croatian wordnet. All three files contain the same data. Two are of the XML type, but formatted according to different schema, and the third is in the JSON format.

Collaborators on CroWN development were: Ida Raffaelli, Krešimir Šojat, Daniela Katunar, Matea Srebačić, Vanja Štefanec, Ana Agić, Daša Berović, Lejla Čolić, Marko Tadić, Božo Bekavac, Željko Agić, Igor Marko Gligorić, Ana Ban.

3.21. Croatian ACD (hrACD)

The following work has been carried out for the tool within CESAR:

- preparation of the hrWaC for automatic collocation extraction
- submitting of hrWaC data for the processing to Lexicom Ltd.
- filtering the results of processing hrWaC for systemic errors
- compilation of metadata and description

The Croatian Automatic Collocations Dictionary is a list of collocations based on hrWaC and created by Lexical Computing Ltd. following the same methodology for all CESAR languages. Since in some parts of hrWaC were poorly tokenised due to the text-encoding exceptions, the original hrACD obtained from Lexical Computing Ltd. was filtered to allow only real words in collocations.

3.22. Croatian Weather Dialogue Corpus (CWEDIC)

The following work has been carried out for the tool within CESAR:

- contact with the out-of-consortium resource provider
- negotiations with the resource provider
- compilation of metadata and description

The Croatian Weather Dialogue Corpus (CWEDIC) contains recordings of prepared dialogue questions by multiple male and female native Croatian speakers for use in weather information dialogue system. This is the first Croatian speech corpus.

3.23. CESAR Aligned Wikipedia Headwords List (hrACD)

The following work has been carried out for the tool within CESAR:

- Wikipedia dumps were collected
- all headwords processed (extracted) and interlingually linked (aligned)
- compilation of metadata and description

The 762,662 entries of the lexicon were built from the Wikipedia dumps of the six CESAR languages by using article titles and their interlingual links to English. In the first phase one lexicon for each CESAR language was built after which those lexicons are merged by grouping together all entries that are connected by interlingual links. If more than one article of a language is connected to a group of articles in other languages (which are actually errors in the structure of the Wikipedias), all article titles are retained, divided by a semicolon. An example of such an entry is sr: "Астеци; Империја Астека". In the final phase category information from the English Wikipedia is added with categories divided by semicolons, and for each non-English entry the number of links to that page in the Wikipedia of the respective language is given.

3.24. Croatian and Slovene NERC models for Stanford NERC (hrStanfordNERC, slStanfordNERC)

The following work has been carried out for the tool within CESAR:

- a Croatian corpus of ca 700,000 tokens compiled and automatically NERC annotated with the rule-based system
- the annotated corpus was manually corrected
- a Slovene corpus of ca 350,000 tokens compiled and automatically NERC annotated
- the annotated corpus was manually corrected
- StanfordNERC system was trained for Croatian and Slovene using the prepared corpora
- compilation of metadata and description

Stanford NER model for named entity recognition and classification (NERC) in Croatian texts is built by using the Stanford Named Entity Recognizer tool (URL http://nlp.stanford.edu/software/CRF-NER.shtml) based on Conditional Random Fields (CRF). The model was trained on a portion of texts crawled from the Vjesnik news portal and manually annotated for ENAMEX TYPE={LOCATION, ORGANIZATION, PERSON}. The manually tagged portion of the text consists of 200.006 tokens in 7.358 sentences, containing 5.966 person tokens, 6.897 organization tokens and 4.784 location tokens. Tokens and named entity tags were used as features in the training procedure. Stanford NER model for named entity recognition and classification (NERC) in Slovene texts is built by using the Stanford Named Entity Recognizer tool (URL http://nlp.stanford.edu/software/CRF-NER.shtml) based on Conditional Random Fields (CRF). The model was trained on a portion of texts selected from the SSJ-500k corpus of Slovene (URL http://www.slovenscina.eu/tehnologije/ucni-korpus) and manually annotated for ENAMEX TYPE={LOCATION, ORGANIZATION, PERSON, MISC}. The manually tagged portion of the text consists of 216.011 tokens in 9.663 sentences, containing 4.204 person tokens, 2.526 organization tokens, 2.421 location tokens and 1.143 miscellaneous tokens. Manually assigned POS/MSD tags and lemmas for the 216.011 tokens were extracted from the corpus and used as features in the training procedure.

3.25. Coral Corpus Aligner (Coral)

The following work has been carried out for the tool within CESAR:

- contact with the out-of-consortium resource provider
- negotiations with the resource provider
- compilation of metadata and description

Coral (CORpus ALigner) is a tool for easy bilingual parallel corpora alignment. It allows both automatical and manual alignment. The main features of Coral are: (1) Automatic segmentation of texts into sentences, (2) manual sentence segmentation editing, (3) automatic parallel text alignment using either the Gale-Church alignment method or a naïve

one-on-one alignment approach, (4) an extremely easy to use manual sentence alignment user interface, (5) exports alignment results into a standard TMX file, (6) runs on all operating systems that can run the Java Virtual Machine, (7) easy installation (a .zip file is simply unpacked onto the user's machine).

3.26. Croatian Sentiment Lexicon (CroSentiLex)

The following work has been carried out for the tool within CESAR:

- contact with the out-of-consortium resource provider
- negotiations with the resource provider
- compilation of metadata and description

CroSentiLex is a sentiment lexicon for Croatian. CroSentilex consists of two files, each containing 37K Croatian lemmas ranked by positivity and negativity, respectively, with the corresponding PageRank scores. The rankings were created automatically based on small positive and negative seed sets and co-occurrence frequencies, using the PageRank algorithm. In addition to the automatically extracted lexicon, human (gold-standard) sentiment annotations for 1200 Croatian lemmas are provided.

4. IPIPAN resources

4.1. Polish Sejm Corpus

The Polish Sejm Corpus contains annotated utterances of Polish Sejm members from terms of office 1-6 (years 1991-2011).

The new release contains corrections of annotation using the latest versions of language tools.

4.2. PoliMorf Inflectional Dictionary

PoliMorf is a morphological dictionary of Polish created from the merger of the two most important competitive morphological resources for Polish – Morfeusz SGJP and Morfologik.

The new version of PoliMorf contains new portion of manually verified data.

4.3. Polish WordNet

The plWordNet (Słowosieć) is a semantic network which reflects the Polish lexical system. The version 2.0 of plWordNet contains new portion of data created by semi-automatic extension of previous version and a portion aligned with Princeton WordNet.

4.4. Nerf – Named Entity Recognition Tool

Nerf is a statistical tool for Named Entity Recognition (NER) based on the Conditional Random Fields (CRF) modelling method. The tool has been constructed as a part of the National Corpus of Polish project. It has been adapted to recognize tree-like structures of NEs (i.e., with recursively embedded NEs) using the Joined Label Tagging (JLT) method. The JLT method is a simple method of encoding NE structures as a sequence of labels. With this method various additional informations about NEs of categorical nature – type, subtype, type of derivation – can be encoded on the level of labels and subsequently recognized using the resultant CRF model. The tool can be configured to use various types of observations during the training and recognition process, for example: lexical informations from textual level, or grammatical informations from morphosyntactic level.

The following work has been carried out for the resource within CESAR since the last delivery:

- Reimplementation of the tool in Haskell has been performed. In particular, statistical back-end has been improved and it now guarantees stability of the training process.
- Implementation has been divided into a collection of packages which can be developed and improved independently. Each package (including the umbrella package nerf) has been uploaded to the Hackage server (http://hackage.haskell.org).
- Support for external dictionaries has been added. Nerf can use one-word named entities and trigger words from external dictionaries to improve results of named entity recognition. The following list of resources is currently supported: PoliMorf

(http://zil.ipipan.waw.pl/PoliMorf), NELexicon (http://nlp.pwr.wroc.pl/en/ tools-and-resources/nelexicon), Gazetteer for Polish Named Entities (http://clip.ipipan.waw.pl/Gazetteer), PNET (http://zil.ipipan.waw.pl/PNET) and Prolexbase (http://zil.ipipan.waw.pl/Prolexbase).

- For memory-efficient dictionary representation a directed acyclic word graph library has been implemented. The library supports incremental dictionary construction and static hashing.

4.13. Morfeusz

Morfeusz is a Polish morphological analyser/synthesizer, currently using PoliMorf lexicon data.

Since delivery 2 (July 2012) new morphological data and inflection patterns have been incorporated into the new delivery of Morfeusz.

4.19. Valence dictionary of Polish

The Polish Valence Dictionary (Walenty) contains a description of argument structures of 1774 Polish verbs and quasi-verbal predicates. The entries are represented through a number of individual frames, each frame corresponding to a set of positions which may be filled by phrases of appropriate types and parameters. Individual positions may be marked for their status as a subject or a passivisable object, and for their role in control relations with other positions in the argument structure.

The Polish Valence Dictionary is an adaptation of the Syntactic Dictionary of Polish Verbs (Świdziński 1994) in a digitised version expanded by Witold Kieraś to include a number of frequent verbs missing from the original dictionary. In addition to expanding the number of frames, the presented resource includes information about new features, including sentential subjects, passivisation, control relations, semantic categories of adverbial phrases, and possibility of coordination of different types of arguments.

The following work has been carried out for the resource within CESAR since previous delivery:

- A new format of a valence dictionary for Polish verbs was established, based primarily on requirements of machine parsers.
- A web application for creating and editing a valence dictionary has been created (Slowal).
- An existing valence dictionary (Marek Świdziński (1994), Syntactic Dictionary of Polish Verbs) was automatically updated to the new format, where automatic convertion was possible, and manually reviewed for corrections.
- The resulting conversion has been released as a preliminary version of the Polish Valence Dictionary under the CC BY-SA licence and as a META-SHARE resource.

- A team of lexicographers was employed for editing the dictionary using Slowal, manually expanding entry structures and adding features established by the new format, missing in the original dictionary. All work was driven by corpus data, with examples corresponding to individual frames stored in Slowal's online database. Annotators were supervised by "superlexicographers", users with elevated status evaluating changes made by lexicographers and correcting mistakes.
- In addition to entries corresponding to the previous version of the dictionary, the valence dictionary was expanded with 335 new entries. New entries constitute the most frequent missing verbs, established through a frequency list generated for the National Corpus of Polish.
- The expanded, manually revised dictionary has been prepared for distribution. The dictionary is a list of valence frames retrieved from the dictionary database in machine-readable text format.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.22. Corpus of the Polish language of the 1960s

Since the last delivery the Corpus of the Polish language of the 1960s was updated after processing it with automated error detection software. Detected errors have been corrected by a linguist.

4.24. Pantera

Pantera is a Brill Tagger for morphologically rich languages.

The following work has been carried out for the resource within CESAR since previous delivery:

- Redesign of the Pantera library API.
- Improvements in sentence segmenter.

4.26. Polish Wikipedia Corpus

Full textual content of Polish Wikipedia as of 28.04.2012, created by applying WikiExtractor to Wikipedia dump and splitting the result into individual articles.

The following work has been carried out for the resource within CESAR:

- Full textual content of Polish Wikipedia (http://pl.wikipedia.org/) has been extracted.
- Stubs, templates, disambiguation pages, history of changes etc. have been removed.
- Resulting articles have been saved independetntly
- Resource description created to maintain META-SHARE compliance.

4.27. SEJFEK4Spejd – a dictionary of Polish economical expressions

SEJFEK4Spejd is the SEJFEK lexicon (Grammatical Lexicon of Polish Economical Phraseology) converted semi-automatically into a lexicalized Spejd shallow grammar. It contains 11,270 automatically generated rules which recognize inflected, case-insensitive multi-word economic terms from the lexicon. Recognized multi-word terms are combined into syntactic words. During the analysis disambiguation (unification and POS-based selection of interpretations) of terms is also performed. Both conversion software and the resulting shallow grammar have been developed within the ERDF Nekst programme.

The following work has been carried out for the resource within CESAR:

- A conversion script has been developed and applied to the SEJFEK lexicon.
- The resulting SEJFEK4Spejd grammar has been proofread.
- An evaluation corpus has been prepared (i.e. manually annotated with economic multi-word terms).
- Both the SEJFEK lexicon and the SEJFEK4Spejd grammar have been evaluated.
- The distribution package for the SEJFEK4Spejd grammar the has been prepared and documented.
- Making the SEJFEK4Spejd grammar available under the CC BY-SA license has been successfully negotiated with the resource creators.
- Making the conversion script available under the GNU GPL v3 license (General Public License version 3) has been successfully negotiated with resource creators.
- The grammar and script metadata description has been created to maintain META-SHARE compliance.

4.28. PNET (Polish Named Entity Triggers)

The Polish Named Entity Triggers (PNET) is an electronic lexicon containing partly inflected external or internal evidences, or trigger words, for Polish named entities (NEs). It counts over 28,000 inflected forms and over 1,500 lemmas. An external or an internal NE evidence is a word or a list of words which appears frequently in the vicinity or inside named entities and is a good indicator of these NEs' types. For instance aktor ('actor') is an external evidence for person names (as in aktor [Zbigniew Buczkowski]), while von is an internal evidence for the same type ([John von Neumann]). Many words can be both external and internal evidences, e.g. jezioro ('lake') is a external evidence in jezioro [Mamry] ('[Mamry] lake') and an internal evidence in Jezioro Białe ('[White Lake]'). External and internal NE evidences can be used in automatic NE recognition via grammar-based or machine-learning methods.

The following work has been carried out for the resource within CESAR:

- The resource has been created semi-automatically from a subset of Polish Wikipedia whose infobox types and categories were manually mapped on the NKJP NE typology.
- The extensional form of the lexicon, containing all inflectional froms of triggers, together with their associated named entity types, has been generated.

- The distribution package has been prepared and documented.
- Making PNET available under the 2-clause BSD licence (FreeBSD) has been successfully negotiated with resource creators.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.29. POLFIE – an LFG Grammar of Polish

POLFIE is an LFG grammar of Polish implemented in the XLE system (Xerox Linguistic Environment), it is being developed within NEKST project. It provides a two-layer representation: constituent structure (c-structure, tree representation) and functional structure (f-structure, AVM representation). It is based on two previous implemented grammars of Polish: its c-structure is based on GFJP2, a DCG grammar used by the parser Świgra, while its f-structure is inspired by FOJP, an HPSG grammar of Polish. Lexical entries used by the grammar are created using Morfeusz, the state-of-the-art morphological analyser for Polish, and converted valence dictionaries used by Świgra.

The following work has been carried out for the resource within CESAR:

- Preparing the first version of a distribution package.
- Making POLFIE available under the GPL (version 3) license.

4.30. Lexeme Forge – a tool for managing the PoliMorf morphological dictionary

Lexeme Forge (Kuźnia) is a web application allowing collaborative creating of new linguistic resources using various imported sources as well as data added directly.

The following work has been carried out for the resource within CESAR:

- A web application for browsing, searching and editing inflection data has been implemented.
- A mechanism for saving changes and an interface for browsing them have been implemented.
- Customizable dictionary exports have been implemented.
- The source code has been released under the 2-clause BSD licence (FreeBSD)
- Project metadata description has been created to maintain META-SHARE compliance

4.31. Slowal – a tool for the managing the Polish valence dictionary

Slowal is a web tool designed for creating valence dictionaries based on the format presented by Filip Skwarski. It describes each entry through a list of individual frames presented as tables which can be expanded by adding new positions, arguments, series of characteristics and examples showing usage of the frame in the Polish language. The users of the tool are divided into following groups:

- Guests who may only view entries and leave comments;
- Lexicographers who are responsible for expanding existing lemma descriptions;
- Superlexicographers who are responsible for checking correctness of the lexicographers' work, managing vocabularies and adding new lemmas.

The following work has been carried out for the resource within CESAR:

- Making Slowal available under the 2-clause BSD licence has been successfully negotiated with resource owners.
- Tool specification has been prepared and approved.
- A first version of a distribution package has been prepared. It contains the source code and installation instructions.
- The tool has been made available for registered users at the http://chopin.ipipan.waw.pl:8087/slowal.
- Instructions for all types of users has been prepared and made available on the official Slowal web site http://zil.ipipan.waw.pl/Slowal.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.32. CorpCor – a tool for detecting errors in corpora

CorpCor is a web-based tool for correcting morphosyntactic annotation in TEI XML encoded corpora (such as NKJP – the National Corpus of Polish). It integrates Poliqarp (http://poliqarp.sourceforge.net/), a library which allows for querying large corpora, with a web-based interface.

The following work has been carried out for the resource within CESAR:

- The application has been used to manually correct morphosyntactic annotation in the 1- million subcorpus of the National Corpus of Polish.
- A first version of a distribution package of the tool has been prepared.
- The code has been released under the GPL v. 3 license.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.33. plWikiEcono – wikipedia-based economical corpus

A corpus of Polish Wikipedia articles from the domain of economy. Annotated automatically (morphosyntactic layer) and converted into TEI format.

The corpus has been created by selecting economy-related categories from the Polish Wikipedia, including economy-related subcategories, downloading all articles (April 2011) from such a list of categories, stripping Wikipedia annotation, tagging the result with TaKIPI 1.8 and converting it to TEI format.

The following work has been carried out for the resource within CESAR:

- A first version of a distribution package of the corpus has been prepared.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.34. plWikiEconoSenses – wikipedia-based economical corpus manually annotated with word senses

A corpus of Polish Wikipedia articles from the domain of economy (plWikiEcono) annotated manually with word senses and converted into TEI format.

The following work has been carried out for the resource within CESAR:

- A first version of a distribution package of the corpus has been prepared.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.35. Prolexbase – a multilingual ontology

Prolexbase 2.0 is a multilingual relational dictionary of proper names, conceived initially at the University of Tours, France and at the University of Belgrade, Serbia, and further developed at the Polish Academy of Sciences (IPIPAN). It contains a language-independent typology of proper names with 4 supertypes and 34 types, as well as various language-independent or language-specific relations (synonymy, meronymy accessibility, variation etc.). A pivot-oriented design of concepts yields alignment of proper names in a language with their counterparts if other languages. A large majority of the data have been extracted from Wikipedia and GeoNames. All data have been manually validated.

Prolexbase creation has been previously supported by the following projects:

- French Technolangue programme from the French Ministry of Industry (2003-2005),
- Egide Pavle-Savicprogramme from the Serbian Ministry of Science, Ministry of Foreign Affairs and the French Ministry of Research,
- ERDF Nekst project.

The following work has been carried out for the resource within CESAR:

- A tool (ProlexFeeder) has been developed in order to populate Prolexbase from open data (mainly Wikipedia and GeoNames), and its usability has been evaluated.
- 40,000 Polish, 33,000 English and 20,000 new French proper names have been extracted from Wikipedia and GeoNames, interlinked, manually validated and inserted in Prolexbase.
- 165,000 inflected forms for Polish names have been automatically generated and manually validated.
- 65,500 language-independent relations have been extracted and manually validated.
- The distribution package for Prolexbase 2.0 the has been prepared and documented.
- Making Prolexbase 2.0 available under the CC-BY-SA license has been successfully negotiated with the resource creators.
- The Prolexbase 2.0 metadata description has been created to maintain META-SHARE compliance.

4.36. Dependency Parsing Model for Polish

Statistical dependency parsing model is trained on the Polish Dependency Bank (PDB, Pol. Składnica zależnościowa) with the publicly available parsing system – MaltParser. MaltParser is a transition-based dependency parser that uses a deterministic parsing algorithm. The deterministic parsing algorithm builds a dependency structure of an input sentence based on transitions (shift-reduce actions) predicted by a classifier. The classifier learns to predict the next transition given training data and the parse history.

The performance of the Polish MaltParser is evaluated with the following metrics: labelled attachment score (LAS – the percentage of tokens that are assigned a correct head and a correct dependency type) and unlabelled attachment score (UAS – the percentage of tokens that are assigned a correct head). Polish MaltParser achieves 84.7% LAS and 90.5% UAS if tested against the PDB validation set and 68.5% LAS/72.2% UAS if tested against the set of 50 manually annotated sentences.

The dependency parsing model for Polish is released under the GNU General Public License v3 (GPL v.3).

The following work has been carried out for the resource within CESAR:

- Formal clarification of the IPR status.
- Resource description created to maintain META-SHARE compliance.

4.37. HateSpeech corpus

HateSpeech corpus in the current version contains over 2000 posts crawled from public Polish web. They represent various types and degrees of offensive language, expressed toward minorities (eg. ethnical, racial). The data were annotated manually.

The following work has been carried out for the resource within CESAR:

- Formal clarification of the IPR status.
- Resource description created to maintain META-SHARE compliance.

4.38. Polish Coreference Corpus

The Polish Coreference Corpus (PL: Polski Korpus Koreferencyjny) is a result of the "Computer-based methods for coreference resolution in Polish texts" project. It contains short fragments (250-350 segments each) of texts randomly selected (preserving the original text type balance) from the full version of the National Corpus of Polish. These fragments are manually annotated with identity coreferential chains and quasi-identity relations. The corpus is supplied in two xml-based formats: MMAX and TEI. It contains automatic morphosyntactic annotation, in TEI format it also has automatic named entity and shallow parsing annotations.

The following work has been carried out for the resource within CESAR:

- Making PCC available under the Creative Commons license has been successfully negotiated with resource owners.
- Corpus metadata description has been created to maintain META-SHARE compliance.
- A first version of a distribution package has been prepared in two formats.

4.39. Polish Coreference Tools

The Polish Coreference Tools is a suite of tools created during the “Computer-based methods for coreference resolution in Polish texts” project. It is going to contain the tool used for manually annotation of the Polish Coreference Corpus and all the automatic coreference resolution tools created during the project. Currently the suite contains the automatic mention detection and coreference resolution rule-based tool, which was used for pre-annotation of the PCC.

The following work has been carried out for the resource within CESAR:

- Making PCT available under the Creative Commons license has been successfully negotiated with resource owners.
- Tools metadata description has been created to maintain META-SHARE compliance.
- A first version of a distribution package of the first tool has been prepared.

4.40. Syntactic-Generative Dictionary of Polish Verbs

Syntactic-Generative Dictionary of Polish Verbs aims at describing the collocative characteristics of Polish verbs. It is "syntactic-generative" because it does not provide full description, but it focuses only at syntactic behavior of verbs while ignoring inflection, word formation and phonology. Semantic information is also reduced to deal only with cases where multiple meanings of given verb implies variation of sentence structure.

The following work has been carried out for the resource within CESAR:

- Development of easier to use, TEI P5 based format for the electronic version of the dictionary.
- Formal clarification of the IPR status.
- Resource description created to maintain META-SHARE compliance.

4.41. Manually aligned CES Polish-English parallel corpus

A corpus of the Centre for Eastern Studies (CES) texts. This resource contains 56 Polish-English texts (6 CES reports, 28 issues of CES studies and 22 issues of the CES publication "Point of View) licensed under the CC-BY-NC license. The texts have been aligned manually on the sentence level using the MemoQ software. The resource is provided as TEI P5-compliant XML files with custom extensions and in the XLIFF and TMX formats.

The following work has been carried out for the resource within CESAR:

- Formal clarification of the IPR status.
- Development of a TEI P5 based format for the corpus.
- Conversion of the corpus into the XLIFF and TMX formats.
- Manual alignment of corpus texts.
- Validation for compliance with the TEI, XLIFF and TMX schemas.
- Resource description created to maintain META-SHARE compliance.

4.42. Polish dictionary for the OpenCYC ontology

The Polish lexicon for OpenCyc has been developed and it contains more than 15,000 mappings between OpenCyc symbols and their Polish names. Besides the mapping, the resource contains morphosyntactic data allowing for the inflection of the Polish lexical units.

The following work has been carried out for the resource within CESAR:

- Making the Polish lexicon for OpenCyc available under Creative Commons – Attribution 3.0 license.
- A set of 24,000 OpenCyc symbols was selected for translation.
- 13,000 OpenCyc symbols were translated into Polish (the remaining part turned out to be hard to translate, e.g. hundreds of names of species found only in North America).
- The source code of the core of the application used to create the translation was publicly released.
- The mapping in the form of a TSV- and YAML-encoded files has been prepared.
- Lexicon metadata compliant with the META-SHARE standard has been created.

4.43. TAG grammar of Polish

A TAG (Tree Adjoining Grammar) has been automatically extracted from Składnica constituency treebank (http://zil.ipipan.waw.pl/Sk%C5%82adnica). The grammar and lexicon are in XMG (http://wiki.loria.fr/wiki/XMG/Documentation) and LEX2ALL (http://wiki.loria.fr/wiki/LEX2ALL) formats respectively. The grammar contains 2802 elementary tree families (1825 initial trees and 977 auxiliary trees). The lexicon contains 11515 lexemes, anchoring a total of 23399 trees (one lexeme can serve as a lexical anchor to more than one tree, e.g. in case of verbs with more than one possible valence frame).

The following work has been carried out for the resource within CESAR:

- The grammar was made available under GPL licence.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.44. Anotatornia – a tool for annotation of corpora

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.

The following work has been carried out for the resource within CESAR:

- Formal clarification of the IPR status.
- Resource description created to maintain META-SHARE compliance.

4.45. Concraft – Constrained Conditional Random Fields Tagging Tool

Concraft is a statistical tool for morphosyntactic disambiguation developed as a part of the CESAR project. It is based on conditional random fields (CRFs) extended with additional, position-wise restrictions on the output domain, which are used to impose consistency between the modeled label sequences and morphosyntactic analysis results both at the level of decoding and, more importantly, in parameters estimation process. The problem of morphosyntactic disambiguation is decomposed into two consecutive stages of the context-sensitive morphosyntactic guessing and the disambiguation proper. The tool is currently adapted to the Polish language and resources, but the method and the library should be applicable to at least other highly inflected languages.

The following work has been carried out for the resource within CESAR:

- Development of the context-sensitive morphosyntactic guessing method and its implementation.
- Evaluation of the tool on the National Corpus of Polish.

- Release of Concraft under the 2-clause BSD license (FreeBSD). Concraft consitutes a collection of Haskell packages and each package (including the umbrella package concraft) has been uploaded to the Hackage server (http://hackage.haskell.org).
- Concraft, its evaluation and comparison to other taggers for Polish has been presented on the COLING conference in a form of the long paper.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.46. Multiservice – a Web service providing linguistic processing of Polish

Multiservice is a web service integration platform for Polish linguisting resources. It is designed for chaining execution of linguistic tools. Processing is triggered by request sent to the web service. Requests are enqueued and handled in asynchronous manner. It is accessible through portable and simple SOAP-based API. New resources can be plugged-in relatively easily using unified API based on Apache Thrift framework. Online demo providing graphical representation of request results is also included.

The following work has been carried out for the resource within CESAR:

- New version of online demo application featuring visualization of results.
- Redesign of service internal architecture.
- Apache Thrift based API used to plug-in new language tools.
- Formal clarification of the IPR status.
- Resource description created to maintain META-SHARE compliance.

4.47. Classification of the DBpedia resources into the OpenCyc taxonomy

The files contain the classification of the Wikipedia articles into the taxonomy of OpenCyc. They were obtained from Wikipedia the infoboxes, introductory sentences, categories and direct mapping between Wikipedia and Cyc. The results were cross-validated and conflicts were resolved using the Cyc built-in inconsistency detection mechanism. Only the most specific types are provided. All more general types might be retrieved using Cyc's built-in taxonomy traversal methods (all-genls function in particular).

The resource has been created outside CESAR and included into META-SHARE by providing a compatible resource description.

4.48. DBPediaExtender

The DBPediaExtender is an information extraction system that extends an existing ontology of geographical entities by extracting information from text. The system uses distant supervision learning – training data is constructed based on matches between values from a knowledge base (DBPedia) and Wikipedia articles.

The system was run on the Polish versions of DBPedia and Wikipedia and extracted more than 44 thousand RDF triples expressing relations between geographic entities from Polish Wikipedia.
Both the DBPediaExtender source code and generated triples are available under the GPLv3 licence.

The following work has been carried out for the resource within CESAR:

- Development of a distant supervision learning algorithm.
- Evaluation of the algorithm with manually created data.
- Extraction of several semantic relations from Polish Wikipedia, which resulted in creation of 44174 RDF triples.
- Resource description has been created to maintain META-SHARE compliance.

4.49. LFG treebank of Polish

A pilot version of the broad-coverage LFG grammar for Polish, constructed within the NEKST project, has been applied to three sets of sentences: two testbanks (FOJP, 333 sentences; and GFJP2, 146 sentences), and a collection of sentences from the National Corpus of Polish (358 sentences), resulting in three collections of parses represented as constituent structure (c-structure) trees and functional structure (f-structure) attribute-value matrices.

The following work has been carried out within CESAR:

- Four LFG-based treebanks of Polish have been made available to INESS for discriminant analysis with the LFG Parsebanker.
- Using the LFG Parsebanker tool developed by colleagues at the University of Bergen, Norway, the three sets of Polish LFG parses were analysed for linguistic discriminants which capture the distinctions between the possible parses and relate linguistic properties to words in the string.
- With the help of f-structure discriminants, the two treebanks containing the test suites of sentences and the treebank of sentences from the National Corpus of Polish have been disambiguated manually, obtaining the correct parse per any parsed sentence. The resulting set of LFG treebanks is the first quality-controlled bank of Polish sentences created with the LFG grammar.

4.50. NKJP1mEcono

Economy-related subcorpus of the National Corpus of Polish, containing manually created sense annotation layer.
The corpus has been created by selecting economy-related paragraphs from the 1 million subcorpus of the National Corpus of Polish. The selection has been made based on the

existence of word n-grams, specific to the domain of economy. The corpus has been manually annotated with word senses of 52 economy-related lexemes.

The following work has been carried out for the resource within CESAR:

- A first version of a distribution package of the corpus has been prepared.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.51. gpwEcono

A collection of stock market reports with manual annotation on the word sense layer and automatic morphosyntactic annotation.

The corpus has been created by downloading stock market reports and stripping HTML markup. Automatic morphosyntactic annotation has been created using the TaKIPI tagger, while manual word sense annotation by linguists using AnotEk.

The following work has been carried out for the resource within CESAR:

- A first version of a distribution package of the corpus has been prepared.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.52. SummaryAnnotationTools

A set of 4 similar desktop applications for summary annotation. These 4 applications facilitate manual annotation of clauses, extract (unconstrained) summaries, extract clause-grained summaries and abstract summaries.
The following work has been carried out for the resource within CESAR:

- Making SummaryAnnotationTools available under the Creative Commons license has been successfully negotiated with resource owners.
- Tools metadata description has been created to maintain META-SHARE compliance.
- A first version of the tools has been published in open access code repository.

4.53. DistSys

A distribution system for texts for any kind of manual annotation. Facilitates random assignment of texts to the annotators, distribution of these texts and finally allows to gather them to form a manually annotated corpus.

The following work has been carried out for the resource within CESAR:

- Tool metadata description has been created to maintain META-SHARE compliance.
- A first version of the tool has been published in open access code repository.

4.54. The Polish SRL corpus

The corpus for evaluation of various unsupervised techniques of semantic roles annotation. The corpus consist of parts of multilingual parallel corpus PELCRA annotated manually with FrameNet-based semantic roles. PELCRA corpus was chosen as a source corpus in order to be able to evaluate techniques of cross-language projection of semantic roles. This is the also the reason why English frames from the FrameNet project were chosen instead of creating their own version designed for Polish.

The following work has been carried out for the resource within CESAR:

- A first version of a distribution package of the corpus has been prepared.
- Formal clarification of the IPR status.
- Corpus metadata description created to maintain META-SHARE compliance.

4.55. Składnica – a treebank of Polish

Składnica is a constituency and dependency treebank of Polish based on sentences from the Polish National Corpus (NKJP), parsed with the Świgra parser and manually verified.

The resource has been created outside CESAR and included into META-SHARE by providing a compatible resource description.

4.56. Świgra – a DCG parser of Polish

Świgra is a parser of Polish using a DCG grammar derived from Marek Świdziński's grammar GFJP (Gramatyka formalna języka polskiego, 1992). Current version includes extensions developed for the Składnica treebank.

The resource has been created outside CESAR and included into META-SHARE by providing a compatible resource description.

4.57. NKJP Model for TnT Tagger for Polish

TnT Tagger model used to tag tokenized Polish text. It was trained on the manually annotated part of the National Corpus of Polish.

The following work has been carried out for the resource within CESAR:

- A first version of a distribution package of the corpus has been prepared.
- The TnT tagger was trained using the manually annotated 1-million-token subcorpus of the National Corpus of Polish, and validated using the usual 10-fold cross-validation method.
- Corpus metadata description created to maintain META-SHARE compliance.

4.58. Polish Automatic Collocations Dictionary

The Polish Automatic Collocations Dictionary contains 30,000 most frequent Polish headwords (nouns, verbs and adjectives), the major grammatical relations that the headword occurs in, and up to 50 collocates it occurs within that grammatical relation.

The following work has been carried out for the resource within CESAR:

- A first version of a distribution package of the corpus prepared by Lexical Computing Ltd.
- Corpus metadata description created to maintain META-SHARE compliance.

4.59. Polish Corpus of Wrocław University of Technology

Polish Corpus of Wrocław University of Technology (PL: Korpus Języka Polskiego Politechniki Wrocławskiej, KPWr) is a corpus of written and spoken documents available on the Creative Common license. It comprises of 350k tokens. The texts are divided into 14 categories (blogs, science, stenographic recordings, dialogue, contemporary and old prose, law, long and short press articles, popular science and textbooks, Wikipedia, religion, official and technical texts). The documents are annotated on the level of chunks and selected predicate-argument relations, named entities, relations between named entities, anaphora relations and word senses.

The following work has been carried out for the resource within CESAR:

- A first version of a distribution package of the corpus prepared by the Wrocław University of Technology.
- Corpus metadata description created to maintain META-SHARE compliance.

5. Ulodz resources

5.1. PELCRA parallel corpus collection

General actions for the parallel corpora resources

- Development and adaptation of annotation standards. A TEI P5 compliant schema was developed for the encoding of parallel corpora.
- Development of a central database for parallel data used to store bibliographic, structural and alignment information designed to handle multiple alignments for the same collection.
- Development of web crawlers, parsers and converters for the acquired data.
- Manual and automatic alignment of the corpora.
- Conversion to TEI P5 and XLiFF formats.
- Documentation and META-SHARE XML metadata headers.
- See: http://pezik.pl/wp-content/uploads/2011/11/LTC_PARALLEL.pdf for further details.

Resource-specific work effort for the respective parallel corpora resources is detailed in sections 5.1.1 -- 5.1.5 below:

- English-Polish CC-BY parallel corpora
- Academia parallel corpus
- Multilingual (Polish-*) parallel corpora
- OSW Polish-English corpus
- PELCRA parallel corpus of literary works

5.1.1. English-Polish CC-BY parallel corpora

This resource, published in Batch 1, contains English-Polish text pairs from the CORDIS news database (10 268 texts), the RAPID press releases of the EU (4 740 texts) and the JRC-Acquis (23319 texts). The articles were web-crawled at ULodz using a custom-build web-crawler and aligned automatically at the sentence level with mALIGNa (Jassem and Lipski 2008). The corpus was further annotated with bibliographic information and made freely available in the TEI P5 and XLiFF formats under the Creative Commons Attribution license (CC-BY).

5.1.2. Academia parallel corpus

This resource consists of 257 texts from “Academia — The Magazine of the Polish Academy of Sciences” aligned manually on the sentence level. Original texts were downloaded from the PAS Academia website. A thorough manual alignment methodology was developed to represent all non-trivial translation equivalence types (e.g. translator insertions, deletions, paraphrases, mergers and compressions, see). The texts are provided as TEI P5-compliant

XML files with custom PELCRA extensions to mark complex translation equivalence types and in the XLIFF format. The corpus has a formal IPR clearance; the license was negotiated with the publisher, written permission was obtained to make it available under the CC-BY-NC license.

5.1.3. Multilingual (Polish-) parallel corpora*

The collection consists of news releases the Community Research and Development Information Centre published at http://cordis.europa.eu/, press releases of the EU available through the RAPID database at http://europa.eu/rapid, European Parliament news available at htpp://europarl.europa.eu and press releases of the European Southern Observatory published at http://www.eso.org. The texts were web-crawled from the respective websites using a set of dedicated web-crawlers developed by the ULodz team. Polish texts and their equivalents in all other languages available were subsequently imported to a central relational database for further processing. The CORDIS collection contains over 24 million words in 6 languages. The ESO collection contains over 1.8 million words in 17 languages. The EuroParl collection contains over 31 million words in 22 languages. The RAPID collection contains over 84 million words in 30 languages. The web-crawled collections were parsed for contents and aligned automatically at the sentence level with mALIGNa (Jassem and Lipski 2008). The resource has been further enriched with structural and bibliographic annotation adhering to the TEI format. An XLiFF version of the data has also been made available. Released in the META-SHARE repository under the Creative Commons Attribution license.

5.1.4. OSW Polish-English corpus

The OSW Corpus is a new parallel resource for Polish acquired within the CESAR project. It contains over 1.4 words of articles published in Polish and English by the *Centre for Eastern Studies*. The articles were web-crawled at ULodz using a custom-build web-crawler and aligned automatically at the sentence level with mALIGNa (Jassem and Lipski 2008). The corpus was further annotated with bibliographic information and made freely available in the TEI P5 and XLiFF formats under the Creative Commons Attribution Non-Commercial license (CC-BY-NC). The OSW Corpus has a formal IPR clearance; the license was negotiated with its publisher and a written permission was obtained on 7th of February 2012 to make it available under CC-BY-NC.

5.1.5. PELCRA parallel corpus of literary works

This resource contains 15 public-domain literary works and their English-Polish/Polish-English translations. The texts were downloaded from repositories of public domain literary works (http://gutenberg.org, http://wikisource.org, and http://wolnelektury.pl), converted to a plain text format and pre-sentencized using the memoQ CAT tool. The texts were subsequently aligned manually on the sentence level by trained annotators. A thorough alignment methodology was used to represent non-trivial translation equivalence types (e.g. translator insertions, deletions, paraphrases, mergers and compressions). The texts are

provided as TEI P5-compliant XML files with custom PELCRA extensions to mark complex translation equivalence types and in the XLIFF format. The corpus has been made available under the Creative Commons Attribution license.

5.2. PELCRA Polish spoken corpus (CC-BY-NC)

The corpus is the largest collection of transcriptions of naturally occurring conversational Polish, compiled by the PELCRA team at the University of Łódź since 2000, initially as part of the PELCRA reference Corpus and later within the National Corpus of Polish. In total, it contains around 1.4 million words of transcriptions of conversations recorded in an informal setting, often without some of the speakers knowing they were being taped (although they had been informed about and agreed to the possibility of being recorded and later granted their permission to transcribe the recordings). Within CESAR the data, previously accessible only through a web interface, has been anonymized, converted and made available in the TEI P5 format.

5.3. PELCRA time-aligned conversational spoken corpus of Polish

This resource is a large multimodal subset of the PELCRA Polish spoken corpus enhanced and made publicly available for the first time in the CESAR project under the CC-BY-NC license by the University of Łódź. The resource contains 73 transcriptions, 368 thousand words, over 43 hours of transcriptions of spontaneous conversations in Polish recorded in informal settings between the years 2008-2010, annotated structurally and bibliographically. The recordings have been manually transcribed orthographically by trained annotators and time-aligned on the utterance level.

The general work on this resource was aimed at making it suitable for release and re-use in the META-SHARE repository and it included:

- Manual alignment of the original recordings with the transcriptions
- Enhancement of a TEI P5-compliant schema for spoken transcripts.
- Development of a central RDB system used to store, process and manage the transcriptions.
- Conversion from temporary formats to the RDB system.
- Anonymization of conversational transcripts, manual correction and completion of the spoken transcripts metadata.
- Documenting the annotation schema (http://pelcra.pl/resources/cesar_header.xml).
- Export to the TEI P5 format.
- Preparation of META-SHARE metadata descriptions, submission to the repository under CC-BY-NC.
- See: http://pezik.pl/wp-content/uploads/2011/11/LTC_PARALLEL.pdf for further details.

5.4. PELCRA word aligned English-Polish parallel corpora

General actions for the word aligned parallel corpora:

- Development and adaptation of annotation standards. A TEI P5 compliant schema was developed for the encoding of word aligned parallel corpora.
- Development of a central database for word aligned parallel data used to store information and perform searches of word and collocation alignments.
- Development of parsers and converters for the acquired data.
- Statistical word level alignment of the corpora.
- Conversion to TEI P5 format.
- Export to database.
- Preparation of documentation and META-SHARE XML metadata headers.

Word aligned English-Polish corpora derived from the sentence aligned corpora (see 5.1). Work on this resource included the development of parsers and converters into a format compliant with GIZA++ word alignment software. Texts were parsed for errors, special care was taken to ensure to eliminate sentence level alignments. Tools were developed to export the raw GIZA++ outputs to TEI and relational database formats.

5.5. PELCRA EN Lemmatizer

A lemmatization dictionary for English texts, which can be used for the alignment and cross-linking of Polish-English resources and development of cross-lingual information-retrieval systems. The creation of the lemmatizer comprised of four major steps. Firstly, a programmatic interface for the relational database containing the BNC was developed, which enabled the extraction of word forms and lemmas and part-of-speech tags. The raw data were parsed for errors. The next step involved the compilation of a deterministic finite automata-based dictionary, using the tools provided with the Morfologik library. Finally, tests were performed to eliminate errors found in the BNC data and to judge the performance of the dictionary with various corpora.

5.6. PELCRA ECL Dictionaries

The PELCRA ECL Dictionaries are a set of Wikipedia-derived thematic Polish-English dictionaries of potential use in NLP applications such as dictionary-based named entity recognition. A programmatic tool for accessing parsing Wikipedia category information was first developed for 11 domains. After extraction, the dictionaries were cross-linked between the two languages and exported to RDF formats. Finally, the generated dictionaries were manually checked for erroneous entries and validated.

5.7. PELCRA Language detectors

The PELCRA language detector is a Java tool for detecting the language of an arbitrary stretch of text. It supports binary classification scenarios which enable the user to detect one of two possible languages, and includes models for distinguishing between Polish and

English (woth possible extensions for other languages). The preparation of the tool involved a number of steps. In order to build the models, training data were generated from the British National Corpus and the National Corpus of Polish (http://nkjp.pl) (10 000 sentences were extracted from each corpus). The training data served as a basis for the generation of ngrams imported into a machine-learning environment. A Java Application Programming Interface (API) was developed to use the models for language detection. Unit and integration tests were performed. Final steps included the preparation of documentation, the setup of a webpage for the resource, and the preparation of Meta-Share headers.

5.8. WebLign website crawler

The WebLign crawler is a custom-built tool used by the PELCRA team to acquire multilingual data. The preparation of the resource involved a number of actions. Firstly, a Java API was developed to provide an infrastructure for creating site-specific website crawlers and HTML parsers. Three customised crawlers and parsers are also delivered for acquiring multilingual texts from the Centre for Eastern Studies (http://osw.waw.pl), the European Southern Observatory (http://eso.org) and the European Parliament (http://europarl.europa.eu) websites. The development of these tools required first an analysis of the structure of the respective websites to extract the relevant data and metadata, and then creating sets of rules for parsing the HTML source of the texts to acquire only the relevant contents. Additionally, a relational database schema was developed to store the results and is delivered along with the tool. The source code and the binary of WebLign has been made available under an MIT license.

5.9. Spelling and NUmbers Voice database

SNUV (Spelling and NUmbers Voice database) is a spelling and number and recognition speech database containing over 220 hours of recordings of Polish speakers reading numbers and spelling words. SNUV was acquired and developed using a controlled crowd-sourcing approach in cooperation with VOICELAB (a Polish voice recognition company). A special spoken data collection platform was developed and set up at the snuv.pl website. Over 400 participants were recruited and asked to produce samples of read speech over a dedicated web application. Several quality checks were run on the collected data, metadata indicating speakers' identities, gender and age were collected and encoded in the database. The database was then processed and made available for release in the form of a single, downloadable bundle containing the audio files aligned with the transcriptions and a relational database dump with metadata descriptions. Considerable logistic efforts were made to complete the SNUV collection data process; the first 300 volunteers received pendrives as rewards for their participation. The database was made available under the liberal CC-BY licence, which makes it a free-of-charge, commercially exploitable resource.

5.10. HASK collocation dictionary (English)

The HASK EN dictionary is a phraseological database used by linguists, language teachers, lexicographers, language materials developers, translators and other language professionals and casual dictionary users. It contains over 150 000 word entries and over 2.8 million combinations Within CESAR, the database was substantially extended. A graphical user interface was developed supporting advanced data exploration and visualisation. Additionally, an web programmatic interface to the HASK dictionary was developed, documented and tested within a number of different scenarios (such as external client applications and MS Word plugins). Available under CC-BY-NC at pelcra.pl/hask_en.

5.11. HASK collocation dictionary (Polish)

The HASK PL dictionary is a phraseological database used by linguists, language teachers, lexicographers, language materials developers, translators and other language professionals and casual dictionary users. It contains over 100 000 word entries and over 5 million combinations Within CESAR the database was substantially extended, a graphical user interface was developed supporting advanced data exploration and visualisation. Additionally, an HTTP API to the HASK dictionary was developed, documented and tested within a number of different scenarios (such as external client applications and MS Word plugins). Available at pelcra.pl/hask_en.

5.12. PELCRA Spoken Learner English Corpus

The PLEC corpus consists of recordings (15 hours) time-aligned with transcriptions (131K words) of conversations (in English) with native speakers of Polish. Learners of different levels are represented in this database. Within CESAR the corpus was annotated for word mispronunciations, quality-checked, time-aligned at the level of utterances and pronunciation error instances, converted to TEI P5 and ELAN formats and made available as a downloadable resource under a CC-BY-NC license.

5.13. Polish-Russian Parallel Corpus

A *manually* aligned Polish-Russian parallel corpus of 4 250 000 words from 20 Polish literary works and 1 legal text and 14 Russian literary texts translated into Russian and Polish respectively. The texts were pre-processed using dedicated alignment software. All corrections to the segmentation and alignment of the translation with the original were performed manually. The texts are provided as TEI P5-compliant XML files with custom extensions to mark complex translation equivalence types, and in the XLIFF and TMX formats. The corpus was originally acquired from the University of Warsaw and converted/enhanced by University of Lodz and the Institute of Computer Science PAS

6. UBG resources

6.1. Serbian Wordnet

This resource, delivered in Batch 1 and upgraded in Batch 2 has been further upgraded for Batch 3. The following amendments were made:

- Serbian Wordnet was enhanced by 814 new synsets, so that this new version now has 18,366 synsets. A considerable part of added synsets belong to Wordnet Affect - synsets corresponding to six emotions: anger, disgust, fear, joy, sadness, surprise. This subset of Serbian Wordnet is now completed. For more details see http://wndomains.fbk.eu/wnaffect.html.
- A number of errors have been corrected, particularly those concerning duplicate literals in two or more synsets.
- The resource is now available for download under license MS NC-NoReD.

6.4. French-Serbian Aligned Corpus

This resource delivered in Batch 1, and upgraded in Batch 2 has been further upgraded for Batch 3. These are the amendments performed:

- Several new literary and newspaper texts were added to the automatically aligned and manually corrected corpus - approximately 100K words for each language. The size of French-Serbian aligned corpus is now 1,948,679 words (1,063,564 in the French part, 885,115 in the Serbian part) and 59,425 aligned segments.
- Resource is available through web interface under license CC-BY-NC.

6.5. Multilingual Edition of Verne's Novel "Around the World in 80 Days"

In this resource, delivered in Batch 1, and upgraded in Batch 2, enhancements have been made for Batch 3:

- The alignment of Slovenian version of the novel was corrected
- The resource is now available for download under license CC-BY-NC.

6.7. English-Serbian Aligned Corpus

This resource, delivered in Batch 2, has further been upgraded for Batch 3. The following amendments were made:

- Several new literary, scientific and newspaper texts were added to the automatically aligned and manually corrected corpus - approximately 320K words for each language. The size of English-Serbian aligned corpus is now 5,078,280 words (2,672,911 in the English part, 2,405,369 in the Serbian part).
- The resource is now available through web interface under license CC-BY-NC.

6.13. Named Entities evaluation corpus for Serbian

Named Entities evaluation corpus for Serbian (SrpNEval) is made available for the first time in Batch 3. It consists of 2000 short news published by Serbian news agencies and daily newspapers in 2005 and 2006. They cover mostly Serbian internal and foreign politics. The size of the corpus is: 2,000 short news, 3,343 sentences, 89,425 words. All texts were automatically tagged with named entities using the system NERanka (see 6.15) and manually corrected resulting in 7,122 named entity tags. Recognized named entities are: persons (full names - 1,648, last names - 152, dignitaries - 24), time expressions (dates - 366 and time of day - 33), measurement expressions - 17, money expressions - 146, amount expressions - 322, percent expressions - 53, geopolitical names (countries - 1,390, cities - 1,440, hydronyms - 11, oronyms - 21), organizations - 1,499. The corpus is available in two variants: Latin alphabet and Cyrillic alphabet.

The resource is available for downloading under license MS NC-NoReD.

6.14. Serbian n-grams

Serbian NGrams (SrpNGrams) represent sets of N-grams extracted from Serbian Lemmatized and PoS Annotated Corpus (SrpLemKor) for N ranging from 1 to 5. SrpLemKor texts consist of: fiction written by Serbian authors in 20th and 21th century, various scientific texts from various domains (both humanities and sciences), legislative texts and general texts.

Each unigram is maximum continuous chunk of non-whitespace lower-case characters. The resource contains all unique N-grams preceded by number of occurrences. It also contains n-gram language models (1-5) in the standard ARPA text and binary format, created by IRST Language Modeling Toolkit. The following results were obtained: 1-grams: 402,397, 2-grams: 2,148,570, 3-grams: 3,309,528, 4-grams: 3,664,760, 5-grams: 3,740,032.

6.15. Named Entity Recognition and Annotation Tool

Named Entity Recognition and Annotation Tool (NERanka) is a web application for named entity (NE) recognition and tagging in Serbian that relies on large-scale lexical resources, such as electronic morphological dictionaries (see 6.9) and a cascade of finite-state transducers. It is based on previously developed components for LeXimir (Tool for lexical resources management and query expansion) and VebRanka (web query expansion tool) and uses various lexical resources. NERanka recognizes several types of named entities, and the user can choose either all NE types or select specific NEs from the Person, Organization, Location, Amount or Time group. Annotation of uploaded text is possible for XML and textual documents. For tagging NERanka uses XML tags which describe the type and subtype of the named entity. The inserted tags are: <pers.spec>, <persName.full>, <persName.first>, <persName.last>, <persName.name>, <pers>, <role>, <persColl>, <roleColl>, <org>, <top.dr>, <top.gr>, <top.hyd>, <top.geo>, <top.reg>, <measure.exact>, <measure.approx>, <measure.range>, <money.exact>, < money.approx>, <money.range>,

<amount.exact>, < amount.approx>, <amount.range>, <percent.exact>, < percent.approx>, <percent.range>, <time.date.abs>, <time.date.period>, <time.date.rel>, <time.hour.abs>, <time.hour.period>, <time.hour.rel>, <time.advdate>, <demonym>.

The resource can be accessed at http://hlt.rgf.bg.ac.rs/VeBranka/NERanka.aspx.

6.16. Serbian Spell Checker

Serbian Spell Checker (SrpSpell) contains three dictionaries (Cyrillic, Latin and combined) and two word lists (Cyrillic and Latin). The combined dictionary is the default one. Words are encoded in UTF-8 code page, normalized to special 8bit code page l-sr (description of this code page can be found in an attached file). GNU aspell files l-sr.cset and l-sr.cmap are included in this package. Support for accented vowels is built in, but in the current release the dictionary does not have any words with accented vowels.

The word list includes 133986 words from SrpKor 2003, the first version of SrpKor (Corpus of Contemporary Serbian). The total size of SrpKor 2003 is approximately 23 million words. SrpKor 2003 is available online (http://korpus.matf.bg.ac.rs) for free, but requests registration. This corpus was also used to check correctness of 284420 words from a corpus

formed from texts published on Internet. Thus, 150080 words are marked as potentially incorrect. In addition, 281865 words, with 762 new words included in release 0.02 are included from the word list of Viktor Kerkez.

SrpSpell word lists were also used to produce MySpell (http://srpski.org/myspell/) and Hunspell (http://gitorious.org/dict-sr) dictionaries for Serbian. MySpell (http://lingucomponent.openoffice.org/MySpell-3.zip) was the former spell checker included with OOo Writer of the free OpenOffice.org office suite, replaced with Hunspell since version 2.0.2 of the OpenOffice.org. Hunspell (http://hunspell.sourceforge.net/) is the current spell checker of LibreOffice, Mozilla Firefox, Thunderbird, Google Chrome, SDL Trados, etc.

The resource is available for downloading under the LGPL license.

6.17. Emotions Annotation Tool

The Emotions Annotation Tools (eEmotion) is a web application for ontological based emotions recognition and tagging of Serbian texts. The application uses RDFS which are created by using nine well-known discrete emotion psychological theories developed by Arnold, Ekman, Frijda, Gray, Izard, Tomkins, Weiner&Graham, Watson and Plutchik. Also, it uses an associative dictionary of Serbian with about 11 thousands words and Serbian morphological electronic dictionary (see 6.09) which contains approximately 4.4 million different inflectional forms of simple words. The application offers a representation of summary results in a graphical form. The application is realized on Csharp Net Framework platform, it is uploaded on http://cvetana.mmiljana.com and can be tested, at this moment, on texts in .html and .txt formats. 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.

6.18. A Web Application for Retrieval and Comparison of NE in Aligned Texts

NERosetta is a multi-user web application that aims to facilitate retrieval and comparison of named entities in a single text or in parallel texts. The main named entity categorization is

realized according to the Quaero annotation recommendation and provides a user with approximately 50 different search options. Each annotated texts can use its own annotation scheme that has to be mapped to the Quaero annotation. Search is performed using Quaero annotation tags, while the user can choose to search for mapped tags, and/or for superordinate and subordinate tags. Registered users have the additional possibility to share annotated resources (in XML format) and annotation schemas for the available set of languages as well as to manage their own resources and schemas. The initial version supports four annotation schemas (Stanford NER 3 and Stanford NER 7 for English, Krstev&Vitas for Serbian and Maurel for French) and three annotated parallel versions of Jules Verne's Around The World in Eighty Days (English-Serbian, French-Serbian and French-English).

The reaource can be accessed at http://arhimed.matf.bg.ac.rs/~andjelka/paralel-extended.

6.19. Rhetorical Figures for Serbian

Rhetorical Figures (SrpRetFig) is a database that consists of 98 rhetorical figures for Serbian related to rehetorical figures for English located at http://rhetfig.appspot.com/list. It is downloadable in xml format for registered users. The SrpRetFig web tool is created for maintaining the database, adding examples in it and sorting by: rhetorical types, linguistic types or linguistic operations.

SrpRetFig was created within the scope of a project aimed at building a domain and application ontologies for rhetorical figures in Serbian as well as tools for annotation of rhetorical figures retrieved from Serbian texts. To this end, the database is used in the process of domain ontology building and in the process of creating an annotation tool.

In the process of building SrpRetFig we collected as many rhetorical figures used in Serbian as could be found in traditional resources. They were then connected conceptually and linguistically to other similar resources for other languages, especially English. The main envisaged usage of this database are various NLP application, but it can be also used for improving human knowledge and for a better understanding of rhetorical figures in Serbian.

7. IPUP resources

NooJ is a linguistic development environment that allows linguists to formalize several levels of linguistic phenomena: typography and spelling; lexicons of simple words, multiword units and discontinuous expressions; inflectional, derivational and productive morphology; local and structural syntax, transformational and semantic analysis and generation.
For each of these levels NooJ provides linguists with one formal framework specifically designed to facilitate the description of each phenomenon, as well as parsing/development/debugging tools designed to be as computationally efficient as possible, from Finite-State machines to Turing machines. This approach distinguishes NooJ from other computational linguistic frameworks which provide a unique formalism based on a compromise between power and efficiency.
As a corpus processing tool, NooJ allows all researchers and professional to extract information from general or technical corpora by applying sophisticated queries based on concepts rather than word forms and build indices, add semantic annotations, perform statistical analyses, etc.

NooJ is one of the most popular linguistic tools, but it was available only on Windows platforms using .NET framework. CESAR project recognized the need of potential NooJ users working on non-Windows platforms and planned to make NooJ platform independent.
This decision is further supported by the following reasons:

- Five of the six languages involved in the CESAR consortium has extensive resources developed in the NooJ linguistic development environment. The tool is also used by the consortium members to perform syntactic analysis in tasks like sentence and clause-segmentation, NP-recognition, predicate identification and the identification of the other sentence constituents (e.g. adverbials).
- NooJ is a truly multilingual platform: resources exist in no less than 14 languages ranging from Arabic, Chinese to Catalan and Hebrew not to mention major languages like Fench, English, Spanish, Portuguese. In other words, members of the partner consortiums have also a vested interest in enhancing and upgrading NooJ.
- Widespread use of NooJ is seriously limited by the fact that although freely available for research purposes it is not open source and suffers from limited interoperability and cross-platform availability.

The work of IPUP team on uploading MONO and Java versions of NooJ in the 3rd batch was actually aligned with the following activities (Task 3.3) as written in the DoW:

- Make NooJ open source
- Make NooJ platform independent by turning the current C# code into Java
- Make NooJ maximally interoperative by making sure it will seamlessly work with major tools

The original idea was to use Java to do this task and this idea was welcome by the IPUP team comprised of experienced Java programmers. However, we remained open for other alternatives (MONO framework) since, as confirmed by our Project Coordinator, the choice of the implementation framework is strictly a technical issue as long the main objectives are satisfied.
Finally, we have implemented both MONO and Java versions of NooJ, while only the Java version is open source.

7.1. MONO Version of NooJ

The MONO version of NooJ runs using the MONO framework on multiple platforms (e.g. Mac OS X, Ubuntu, Linux, etc.).

The activities performed by the IPUP team related to the implementation of MONO version of NooJ are the following:

- Get familiar with C# programming language. The IPUP team is comprised of experienced Java programmers, but with no prior experience with C#. During this activity IPUP team members learned C#.
- Examine the MONO framework. IPUP team members were newcomers to the MONO framework, so within this activity we examined the features and abilities of this framework but also the corresponding development environment.
- Create the working environment. IPUP team configured the version control system to support parallel work on NooJ code.
- Create mock-up prototypes. We have created several simple mock-up prototypes to check if NooJ can be successfully ported to MONO framework.
- Examine the functionality of NooJ. The IPUP team members are not computer linguists, therefore they were familiar neither with computer linguistics nor with NooJ. To be able to implement the MONO version of NooJ it was necessary to examine in detail the functionality offered by NooJ.
- Understand the NooJ C# code. The NooJ C# code is comprised of approximately 90.000 lines of code and IPUP team had to examine and understand this code before we started with the implementation of the MONO version.
- Create open source file conversion solutions. The .Net version of NooJ used the Microsoft proprietary solutions for converting files in different formats (e.g. html, xml, doc, docx, pdf) to the txt format. However, this was not feasible in the MONO version, therefore IPUP team created the appropriate open source solutions.
- Create new dictionary editor. The dictionary editor in the .Net version of NooJ didn't use a strict grammar for dictionary entries and consequently it couldn't facilitate a precise error reporting. The strict grammar for dictionary entries was developed supervised by Max Silberztein as the author of NooJ and then IPUP team developed a new dictionary parser and the corresponding dictionary editor. The new dictionary

editor supports coloring of dictionary entries and precise error reporting. This new dictionary editor was then fully integrated in NooJ.

- Porting NooJ to MONO framework. Microsoft is the owner of many patents related to the implementation of .Net platform, therefore the MONO framework, which supports C#, was actually implemented from scratch and consequently the functionality and behavior of many GUI controls and classes differ from the original .Net controls and classes. Due to these differences, porting NooJ GUI to MONO platform proved to be a difficult task, and required a thorough examination of the complete existing code and performing changes where it was necessary. This activity was the most labor intensive.
- Extensive testing of MONO version of NooJ on multiple platforms. MONO version of NooJ was thoroughly tested by IPUP team on three major platforms: Linux, Mac OSX and Windows. Although we used the same MONO framework, apart from bugs detected on all platforms, some platform specific bugs were also detected.
- Debug MONO version of NooJ on multiple platforms. The majority of bugs detected were related to the MONO implementation and behavior of GUI controls which sometimes differ substantially from the .Net GUI controls. The behavior of MONO GUI controls is not documented anywhere, hence it was extremely difficult to correct these bugs.
- Create installation scripts and installation documentation. It is not enough to install the core MONO framework to be able to start the MONO version of NooJ. IPUP team had to determine which packages are missing and to install them together with NooJ itself for each major platform. Brief installation instructions are also given in the corresponding documentation.

All these activities were performed in close collaboration with Max Silberztein as the author of NooJ. Without his constant support it would be impossible to finish all these activities in one year. The IPUP team of 5 members was involved in the above described activities on the implementation of MONO version of NooJ.

7.2. Java Version of NooJ

Java Version of NooJ runs on Java Runtime Environment available for various platforms including Windows, Max OS X, Linux, Solaris, etc.

The task of implementing Java version of NooJ consisted of the following main activities:

- Decouple the engine from the GUI and define the corresponding API. In the .Net version of NooJ, the NooJ engine was not clearly separated from the GUI and, as an effort to increase the interoperability of NooJ, we have decoupled them and defined the corresponding API.
- Port the NooJ engine to Java. Since Java as a programming language is similar to C#, this was a relatively straightforward task where the problems that emerged were mainly related to the differences between C# and Java regarding: passing parameters

by reference and the *goto* command, which are not supported in Java, a limited number of data types supported for the switch/case command in Java, different behavior of some methods handling strings (e.g. substring) in Java, etc. Some code cleansing has been performed, but also unit testing and documenting.

- Create open source file conversion solutions. The .Net version of NooJ used the Microsoft proprietary solutions for converting files in different formats (e.g. html, xml, doc, docx, pdf) to the txt format. However, this was not feasible in the Java version, therefore our team created appropriate open source solutions.
- Specify new binary file formats used by Java version of NooJ. The binary file formats used in the C# version (nod, nof, nog, not, noc, nom) are translated into new Java binary file formats (jnod, jnof, jnog, jnot, jnoc, jnom).
- Analyse different Java GUI libraries and implement prototypes. We have analyzed several Java GUI libraries in order to find the one which is the most suitable for the implementation of NooJ GUI in Java. Since NooJ C# GUI evolved during several years through the feedback of NooJ users, the main requirement was to create NooJ GUI in Java that will be the most similar to the original GUI.
- Port NooJ's GUI to Java. The whole .Net version of NooJ GUI had to be implemented from scratch using the Swing GUI components, trying to follow the appearance of the original .Net version. Swing has been chosen for two reasons: 1) it supports the Multiple Document Interface (MDI) used in the .Net version of NooJ GUI; 2) its text editor GUI component supports Unicode characters, which is essential for languages like Arabic, Vietnamese, Hebrew, etc.
- Integration of the engine and the GUI. After the NooJ engine and GUI were successfully ported to Java, they had to be properly integrated and thoroughly tested .
- Extensive testing of Java version of NooJ on multiple platforms. MONO version of NooJ was thoroughly tested by IPUP team on three major platforms: Ubuntu, Mac OSX and Windows.
- Debug Java version of NooJ on multiple platforms. The bugs found were mainly related to the implementation of NooJ GUI in Java, but also to differences in C# and Java as programming languages..
- Create installation documentation. Brief installation instructions are given in the corresponding documentation.

In performing all these activities we closely communicated and collaborated with Max Silberztein as the author of NooJ.

8. IBL resources

8.1. Bulgarian National Corpus

The work carried out on the Bulgarian National Corpus (BulNC) consists of several parts – corpus expansion, annotation, structure and metadata enhancement and update of information. The actions carried out within the CESAR project cover the followings:

- Corpus expansion (as of January 2013, its core Bulgarian part consists of over 1.2 billion words and the overall size is over 5.4 billion words)
- Compilation of metadata and description of the newly added texts
- Standardisation of text samples and corpus structure incl. the newly added texts
- Linguistic preprocessing and annotation of the newly added texts
- Maintenance and deployment of programming tools for corpus compilation and annotation
- Upgrade of the web search engine
- Update of the corpus website
- Development of a collocation service
- Conversion of the corpus format
- Clearance of license terms

The enlargement of the BulNC has involved not only amassing of Bulgarian texts, but also compilation of parallel corpora with Bulgarian as a pivot language. The texts in other languages obligatorily have a Bulgarian counterpart in the Bulgarian part of the corpus.
Currently (January 2013), the corpus core consists of over 1.2 billion words in about 240,000 texts. So far 47 foreign languages have been included totaling app. 4.2 billion words. Thus, the overall size of the corpus exceeds 5.4 billion words.
UTF-8 encoding was used for all text samples and texts in other encodings (e. g., Windows-1251) were converted. All text samples are stored in plain text format (.txt) and placed in the relevant directory according to style and subdirectory according to its primary domain. Each text sample is given a unique ID which identifies among in the corpus categories and is also its filename. All newly added texts in BulNC have been automatically lemmatised and morphologically annotated.
All texts were supplied with extensive metadata description compliant with the established standards. The corpus is supplied with three levels of annotation: a detailed metadata description: each text is supplied with editorial (author's name, text title, source, etc.) and classificatory metadata (general category, domain, genre); monolingual annotation: tokenisation, sentence splitting, POS tagging, lemmatisation, word sense annotation; and multilingual annotation: alignment at different levels, currently sentence and clause level.
Corpora for the other languages were tokenised, sentence-split and aligned.

A special corpus search system allowing complex queries for extracting the metadata and compiling the corpus description from the markup formats was finalised. The Corpus Collocations service is a web service for online collocations search and different types of statistics over the BulNC. The Corpus Collocation service employs the free-of-charge NoSketchEngine, a system for corpora processing that combines Manatee and Bonito. The Collocation service is a RESTful webservice, supporting complicated queries through http. The query returns the collocations of a given word in the NoSketchEngine format. The system supports additional arguments, namely all that are accepted by NoSketchEngine, provided with default values.
The information in Bulgarian and English on the webpage (http://www.ibl.bas.bg/en/BGNC_en.htm) was updated.
The corpus is a pseudocorpus - the proper texts cannot be distributed, only small excerpts are available through the query interface. The text excerpts and query results are offered under META-SHARE NoRedistribution Non-Commercial license for free.

8.2. Bulgarian-X Language Parallel Corpus

Bulgarian-X Language Parallel Corpus (Bul-X-Cor) includes parallel corpora of 48 languages – English, German, French, Slavic and Balkan languages, as well as other European and non-European languages. The following work has been carried out for the resource within the CESAR project:

- Collection of text samples (up to the present size of the corpus of app. 4.2 billion tokens as of January 2013)
- Compilation of metadata and text description
- Standardisation of text samples and corpus structure
- Linguistic preprocessing, annotation and alignment
- Maintenance and deployment of programming tools for corpus compilation and annotation
- Enhancing the web search engine for parallel corpora support
- Creating of the corpus website
- Development of the collocation service
- Conversion of the text format
- Update of the query format
- Update of the visualisation format
- Update of the system for dealing with very large data sets
- Clearance of license terms

At present, there are 47 parallel corpora in the BulNC which are collectively named Bulgarian-X Language Parallel Corpus and which contain app. 4.2 billion tokens (during the last six months it has been increased more than twice), comprising the biggest parallel corpus

of Bulgarian. Languages are not equally represented: the largest parallel corpus is the Bulgarian-English parallel corpus (280.8 and 283.1 million words for Bulgarian and English respectively); there are 18 other corpora of over 200 million tokens per language, 2 corpora between 100 and 200 million tokens per language, 11 parallel corpora of size in the range 1-15 million tokens per language, and the rest 15 are below 1 million, with the smallest corpora being Azerbaijani, Armenian, Georgian, Kyrgyz, Tajik (with less than 200,000 tokens) and Japanese corpus (with less than 50,000 tokens). Each parallel subcorpus within Bul-X-Cor mirrors the structure of the BulNC.
Several sets of tools are used for performing various tasks – collection of texts, compiling metadata, linguistic annotation, etc. To ensure easy collaboration, knowledge exchange, code reusability a uniform framework for all programming tools is established – for development and debugging, building test environments, documenting source code, creating repositories of programs.
New information and description of Bul-X-Cor is added to the website of the Bulgarian National Corpus, plus some publications.
The corpus is a pseudocorpus - the proper texts cannot be distributed, only small excerpts are available through the query interface.
Collocations service is a web service for collocations search and different types of statistics over the Bulgarian-X Language Parallel Corpus.The Collocation service employs the free of charge NoSketchEngine, a system for corpora processing that combines Manatee and Bonito. The Collocation service is implemented as a RESTful web service, supporting complicated queries through http. The query returns the collocations of a given word in NoSketchEngine format. The system also supports additional arguments, namely all that are accepted by NoSketchEngine, provided with default values. The service is protected with the HTTP Digest authentication.
New indexing of the Bulgarian-X Language Parallel Corpus is performed. The corpus format is converted to a format readable by the NoSketchEngine indexing machine.
The text excerpts are offered under META-SHARE NoRedistribution Non-Commercial license for free.

8.3. Bulgarian wordnet

The Bulgarian wordnet (BulNet) is a network of lexical-semantic relations, an electronic thesaurus with a structure modelled on that of the Princeton WordNet.
The following work has been carried out for the resource within the CESAR project:

- Enlargement of Bulgarian wordnet with new synsets, literals and relations
- Review and correction of existing synsets, literals and relations
- Automatic validation of data consistency
- Releasing a new version of the resource via ELDA
- Upgrading do the web site of the resource

The Bulgarian wordnet was manually enlarged with around 5,000 new synsets to 49,189 (as of January 2013). The Bulgarian wordnet is enriched as well with new relations, namely, hyponyms, meronyms, antonyms, etc. defined between the newly added and already existing synsets. Extension in the number of literals is not only a consequence from the increasing of the wordnet itself but it is due to some specific peculiarities of Bulgarian - verbal aspect, rich derivational system, etc.
Validation for consistency was carried out automatically followed by bug fixing, and manual correction of errors and inconsistencies. In the process of the database enrichment, spelling and grammar errors in the existing synsets (literals, definitions, notes, examples) were identified and manually corrected. There were also instances of erroneously grouped or missing literals. Those also had to be manually removed, add, merged or split. Automatically generated relations are manually validated and if necessary corrected.
The latest version of the Bulgarian wordnet is spread by ELDA. The resource is offered under META-SHARE NoRedistribution Non-Commercial license for a fee, and under META-SHARE NoRedistribution Commercial license for a fee.

8.4. Lists of Bulgarian Multiword Expressions

Several methods for automatic extraction and classification of Bulgarian Multiword Expressions **(**MWEs) have been developed and tested. The main method consists of the following stages:

- Annotation of the corpus – POS tagging, lemmatisation, chunking.
- Extraction of MWE candidates via syntactic filter – 16 types of syntactic constructions are considered; most frequent are AN (70.4%), NN (15.4%) and NPN (9.1%) (A – adjective; N – noun; P – preposition). Altogether, 2.2 million candidates are registered.
- Frequency analysis and MWE extraction – constructions with higher frequency are extracted.
- Extraction using association measures – candidates showing higher association between candidates are extracted.
- Rule-based classification – a set of rules is assembled to distinguish between types of MWEs.

We adopt the classification of multiword expressions (MWEs) developed by Baldwin et al. (Baldwin, T., C. Bannard, T. Tanaka, D. Widdows. An Empirical Model of Multiword Expression Decomposability. In: Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment. 2003) who distinguish between non-decomposable, idiosyncratically decomposable and simple decomposable MWEs. Further, we divide simple decomposable MWEs into 10 categories based on pragmatic factors – whether they are or contain a named entity (NE). Free collocations are free phrases (non-

MWEs) which are statistically marked (Baldwin, 2004), i.e. appear with high frequency in a corpus, but are not linguistically marked. The lists of Multiword expressions are the result of automatic and semi-automatic tagging and classification of the corpus *Wiki1000+* (13.4 million tokens): *Non-decomposable* - 700, *Idiosyncratically decomposable* - 3,156, *Simple decomposable* (NEs without connection between elements - 36,932, NEs with a meaningful element(s) - 11,248, Non-NEs with a vague connection between components - 1,46, NEs with meaningful components but connection difficult to restore - 1,086, NEs with descriptor and additional element - 18,962, Non-NEs with a NE as one of the components - 27,373, Non-NEs with a standard, easy to restore connection between components- 140,394, NEs with a standard, easy to restore connection between components - 16,653, Non-NEs with explicit connection between components - 1,468), "*Free collocations*" - 49,651, *Free phrases*- 1,197,762.
The lists are distributed under META-SHARE NoRedistribution Non-Commercial license for free.

8.5. Bulgarian Frequency Dictionary

The Bulgarian Frequency Dictionary is based on the monolingual part of the Bulgarian National Corpus and represents words (lemmas) with the frequency of their occurrences in the corpus. The work on Bulgarian Frequency Dictionary has several stages:

- Preparation of the corpus
- POS tagging and lemmatisation
- Word frequency count – in stages

The frequencies are automatically collected and more efficient methods for compilation of frequency lists and dictionaries are still being investigated. The compilation of the frequency dictionary is performed in stages – compilation of the dictionary on smaller parts of the corpus, followed by merging.
The Frequency dictionary is distributed under META-SHARE NoRedistribution Non-Commercial license for free.

8.6. Hydra - tool for developing wordnets

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

The following work has been carried out for the resource within the CESAR project:

- Refactoring the code
- Simplifying the table descriptors
- Fixing bugs
- Developing installation and user manuals
- Developing a web site of the resource
- Clearance of license terms

As a platform-independent system, Hydra has been successfully tested under Linux and Windows. The existing body of code was restructured to simplify its internal structure without changing its external behaviour, undertaken in order to improve some of the nonfunctional attributes of the software.
The source code has been made available under the GPLv3 license.

8.7. Chooser - annotation tool

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

- Refactoring the code
- Fixing bugs
- Developing installation and user manuals
- Developing a web site of the resource
- Clearance of license terms

The source code has been made available under the GPLv3 license.

8.8. Bulgarian Sentence Splitter and Tokenizer

The sentence splitter marks the sentence boundaries and the tokenizer marks string of symbols in raw Bulgarian text.
The following work has been carried out for the resource within the CESAR project:

- Porting to Linux
- Fixing bugs
- Clearance of license terms

The sentence splitter applies regular rules and lexicons. Both - regular rules and lexicons - are manually crafted by an expert. Lists of lexicons (for recognizing abbreviations after which there must be or there might be a capital letter, a number, etc. in the middle of the sentence) are applied before the regular rules. The lexicons are compiled by a separate tool - the Lexicon compiler, as minimal acyclic final state automata which allows an effective processing. Sentence borders are represented as a position and length which allows the incoming text to be kept unchanged as well as an easy integration in different systems for annotation.
The tokenizer demarcates strings of letters, numbers, punctuation marks, special symbols, combinations of them and empty symbols. Regular patterns are used to recognize some simple cases of named entities that mean dates, fractions, emails, internet addresses, abbreviations, etc. The tokenizer classifies each recognized token (for example: small Cyrillic letters, capital Latin letters, etc.). The tokenizer utilizes finite state transducers for token recognition and type matching.
The resource is distributed under META-SHARE NoRedistribution Non-Commercial license for free.

8.9. Web based infrastructure for Bulgarian data processing

A web based infrastructure combines Bulgarian tokenizer, sentence splitter, tagger and lemmatizer.
The Bulgarian POS tagger marks up each word with the most probable Part of Speech and unambiguous morphosyntactic information among the set of tags associated with a given word. The tagger is based on SVM (Support Vector Machines) learning. The tagger predicts the POS tag of a word based on a set of features describing the word and its context. The trained model is applied to disambiguate texts. The precision of the tagger up to the moment is 96,58%.
The tagger exploits the SVMTool, an open source utility for training of tagger models and their application for POS disambiguation. To improve the robustness of the SVMtool an alternative disambiguation module has been developed in C++. The new implementation

provides an integration with the lower levels of annotation, full Unicode support and improves the model loading speed.
The functionalities of the tools can be accessed trough a RESTful web service or by means of asynchronous input/output processing that permits other processing to continue before the transmission has finished. The infrastructure consists of three main components: Frontend, Backend and TaskDispatcher. The Frontend supports the access policies. The Backend combines the Bulgarian language processing tolls as a server application which handles the requests over tcp/ip. The TaskDispatcher manages the asynchronous tasks.
The following work has been carried out for the resource within the CESAR project:

- Implementation of the web based infrastructure
- Fixing bugs
- Clearance of license terms.

The obtained results are distributed under META-SHARE NoRedistribution Non-Commercial license for free.

8.10. Bulgarian Wordnet web access

Wordnet service is an online service that gives the users access to a subset of the Bulgarian wordnet (BulNet).
The following work has been carried out for the resource within the CESAR project:

- Development of the web service
- Conversion of the format
- Clearance of license terms

The system is a RESTful web service that supports two sorts of queries through http. The first one is search for objects where the query described in the WordNet modal language returns a list of object identifiers for which it is true, and the second one searches for information about objects and returns a list of data for: literal (identifier, word, lemma); synset (identifier, ili, POS, definition, stamp, bcs, language (identifier), frequency); and note (identifier, text). The queries support non-obligatory parameter (format) showing the type of the result. If the value of the format is json, the result is coded as json, otherwise it is not coded.
Users can search for synonyms, hypernyms, antonyms, and translation equivalents of different words and lemmas in the following language pairs: English-English, English-Bulgarian, Bulgarian-English, and Bulgarian-Bulgarian.
The results are offered under META-SHARE NoRedistribution Non-Commercial license for free.

8.11. Bulgarian Spell Checker for Windows

The system for automatic spelling checking WinEst for Microsoft Office detects and marks the incorrectly written words in a text and suggests the most probable candidates to correct the errors. WinEst offers the entire potential of the contemporary spelling correction:

proficiently compiled dictionary, which contains over a million and a half words, and replacement suggestions ordered according to their probability.
WinEst is based on the Electronic Grammar Dictionary of Bulgarian, developed at the Department of Computational Linguistics, which contains over 85,000 words. It contains logic for detection of careless mistakes (wrong key pressed, letter swapping, skipped letters or extra letters), identifies errors of ignorance and integrates perfectly into the dictionaries used in Microsoft Office. Its functionality is realized through the use of minimal acyclic deterministic automata and Levenshtein automata allowing maximum speed, precision and coverage.
The following work has been carried out for the resource within the CESAR project:

- Fixing of bugs
- Deployment of the system
- Tests
- Clearance of license terms

WinEst works perfectly both with the standard BDS layout and with the various phonetic layouts. WinEst is a 32-bit module and thus requires a 32-bit Microsoft Office.
The tool is offered under META-SHARE NoRedistribution Non-Commercial license for free.

8.12 Bulgarian Spell Checker for Mac OS

MacEst is a system for automatic spell checking for Mac OS X detects and marks the incorrectly written words in a text and suggests the most probable candidates to correct the errors. MacEst offers proficiently compiled dictionary, which contains over a million and a half words, and replacement suggestions ordered according to their probability. MacEst is based on the Electronic Grammar Dictionary of Bulgarian, developed at the Department of Computational Linguistics at the Institute for Bulgarian Language, which contains over 85,000 words. It contains logic for detection of careless mistakes (wrong key pressed, letter swapping, skipped or extra letters), identifies errors of ignorance, and integrates perfectly into the dictionaries used in Mac OS. The tool functionality is realized through the use of minimal acyclic deterministic automata and Levenshtein automata, which allow maximum speed, precision and coverage.
The following work has been carried out for the resource within the CESAR project:

- Fixing bugs
- Deployment of the system
- Tests
- Clearance of license terms

MacEst is available for all applications that use Mac OS X spell checking system.
The tool is offered under META-SHARE NoRedistribution Non-Commercial license for free.

8.13. Bulgarian Spell Checker Web Service

The Spell Checker is integrated as a web service – both web service integration and online spell checking are possible. The Spell Checker is based on the construction of a dictionary in a minimal acyclic deterministic automaton and offers replacement suggestions on the basis of Levenshtein automata.
WenEst allows the users to check and correct Bulgarian texts on the Internet. The Spell Checker web service can be used in different blogs, chat forums, online shops, media, and everywhere when creating Internet content to assist the correct writing of Bulgarian texts.
The following work has been carried out for the resource within the CESAR project:

- Further development of the web service
- Deployment of the system
- Tests
- Clearance of license terms

The results are offered under META-SHARE NoRedistribution Non-Commercial license for free.

8.14. Dictionary of Synonyms in Bulgarian Language

The Dictionary of Synonyms in Bulgarian Language covers the body of synonyms in modern Bulgarian. It contains ca. 27,000 unique word forms pertaining to four parts-of-speech, distributed into synonym sets, as follows: verbs: 2,137 synonym sets, containing a total number of 10,000 words; nouns: 3,240 synonym sets, containing a total number of over 12,000 words; adjectives: 2,496 synonym sets, containing a total number of over 10,000 words; adverbs: 910 synonym sets, containing a total number of over 3,800 words.
The following work has been carried out for the resource within the CESAR project:

- Compilation of synonym sets of words pertaining to different parts-of-speech (nouns, adjectives, adverbs, verbs).
- Tests, verification and editing of the data.
- Entering the data into a database.
- Clearance of license terms

The dictionary is distributed under META-SHARE NoRedistribution Non-Commercial license for free.

8.15. Dictionary of Antonyms in Bulgarian Language

The Dictionary of Antonyms in Bulgarian Language covers the body of synonyms in modern Bulgarian. It contains about 8,500 unique word forms pertaining to four parts-of-speech, distributed into 3,644 antonym lists, as follows: verbs: 571 antonym lists and a total number

of over 3,000 words; nouns: 1,399 antonym lists and a total number of over 5,000 words; adjectives: 1,092 antonym lists and a total number of over 4,100 words; adverbs: 582 antonym sets and a total number of over 2,100 words.
The following work has been carried out for the resource within the CESAR project:

- Compilation of antonym pairs and lists involving different parts-of-speech (nouns, adjectives, adverbs, verbs).
- Tests.
- Editing of the data.
- Entering the data into a database.
- Clearance of license terms.

The dictionary is distributed under META-SHARE NoRedistribution Non-Commercial license for free.

8.16. Dictionary of Neologisms in Bulgarian Language

The Dictionary of Neologisms in Bulgarian Language contains over 2,200 new words and 160 new multiword units (compounds and terminological units) that have entered the Bulgarian language in the past 20 years. Each entry contains information about: part-of-speech (for lexemes); origin (for borrowed words); stylistic and grammatical notes; lexical meaning; synonyms and antonyms (if available). If necessary, short examples illustrate the use of the neologism in context.

The following work has been carried out for the resource within the CESAR project:

- Selection of new words and expression that were extracted from a frequency list.
- Attribution to part-of-speech, and entering information about origin (for foreign words).
- Interpretation of the meaning.
- Tests.
- Editing of the data.
- Entering the data into a database.
- Clearance of license terms

The dictionary is distributed under META-SHARE NoRedistribution Non-Commercial license for free.

8.17. Register of Phraseologisms in Bulgarian Language

The Register of Phraseologisms in Bulgarian Language covers the body of phraseologisms in Bulgarian language. It contains phraseological units, as defined by and found in the main Bulgarian phraseological dictionaries. The total number of the phraseological units is over

10,000, and the unique words – over 6,400. When searching in the register for a word, a list of all phraseological units containing the searched word (or a part of it) is composed. Each phraseological unit has a reference note to the source from which it was extracted. Thus, the register provides information on lexicographic editions, in which the meaning of the phraseological unit can be found.
The following work has been carried out for the resource within the CESAR project:

- Excerption of phraseological units from various phraseological dictionaries.
- Attribution of information about the source.
- Tests, verification and editing of the data
- Entering the data into a database.
- Clearance of license terms

The register is distributed under META-SHARE NoRedistribution Non-Commercial license for free.

8.18. Corpus of Spoken Bulgarian

The Corpus of Spoken Bulgarian (SpokenBg) is a selection of data of spoken Bulgarian language incl. data from interviews, media and formal speech, student speech, academic speech, colloquial speech. Part of the corpus contains edited versions of transcripts of conversational speech that have been prepared within the CESAR project, including:

- Editing to convert the files from a semi-phonetic transcription to standard orthography with original semi-phonetic transcripts presented together with the edited versions in a paragraph-aligned display.
- Removing some feature related to specific instructions for transcription of the original raw transcripts, incl. metadata relating to speakers, space-consuming coding for non-verbal gestures, as well as empty lines and similar noisy data.
- preparing a parallel paragraph-level alignment to allow for easy comparison of the two versions.

The total size of the corpus is 523,128 signs as of the end of 2012.

8.19. Corpus of Colloquial Bulgarian

The Corpus of Colloquial Bulgarian is a selection of data of oral forms of contemporary Bulgarian language.
The following work has been carried out for the resource within the CESAR project:

- Extending/enlarging the data of spoken Bulgarian with new files – transcription of audio or video files of media (TV and radio programs) and everyday oral communication.
- Producing html documents from raw transcripts (processing new files which were not part of BgSpeech database) – 7 files containing transcripts from 2004; 2 files

containing transcripts from 2005; 13 files containing transcripts from 2009, and 9 files form 2012.
- Development of a conceptual framework for orthographic normalization of raw transcripts and work plan for their processing.
- Normalization of existing transcriptions.
- Manual/visual inspection (check-up) of orthographically edited transcripts, and verification.
- Summary of metadata.

The corpus amounts to 357,584 signs at the end of 2012.

8.20. TREFL – Translation Reference Library

TREFL is a portable, multifunctional database management application for Windows, having the combined characteristics of both a Translation Memory System (bilingual databases, fuzzy matching, concordance, alignment, importing and exporting translation memories, etc.) and those of an Internet/Desktop Search Engine, plus elements of semantic search. It is intended to be used as a simple, versatile, portable, effective and customizable reading, writing and translation aid tool for very large databases management.
The following work has been carried out for the resource within the CESAR project:

- Further development of the tool
- Update of the database
- Clearance of license terms

8.21. SARP- Speech Analyzer Rapid Plot

The SaRP tool is an extension to the programme Speech Analyzer version 3 or later that allows managing databases of spoken language samples and creating informative charts in an easy and interactive manner. It works under Windows. The tool gives computer generated feedback on vowel production by language learners. It is designed for automatic or semi-automatic (interactive) retrieving of formant values, and easily creates, saves and opens vowel charts. SARP offers support for multiple data sets and gives vowel charts comparison by superimposing control charts and user charts; numerical or visual/graphical editing of the charts and quick-commands: create, move, delete, lock/unlock markers. It also calculates and graphically represents the mean values, and has an integrated library of vocal samples.
The following work has been carried out for the resource within the CESAR project:

- Further development of the tool
- Functionalities upgrade
- Clearance of license terms

8.22. RTComp - Real Time Comparison

RTComp is a tool for Windows that 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.
The following work has been carried out for the resource within the CESAR project:

- Further development of the tool
- Finalisation of the first version
- Clearance of license terms

8.23. Wiki1000+ corpus with annotated MWEs

Wiki1000+ is a corpus of articles from Wikipedia, compiled for the purposes of the study of multiword expressions (MWEs) in Bulgarian. The Wiki1000+ corpus contains 6,311 text samples with at least 1000 tokens each, amounting to 13.4 million tokens. The corpus is a part of the Bulgarian National Corpus. Wiki1000+ is annotated with the following linguistic information: sentence boundaries, tokenisation, lemmatisation, POS tagging, and MWE annotation. MWE annotation includes MWE id, labelling the components of the MWE and determining the type of the MWE according to a classification based on idiomaticity.
The following work has been carried out for the resource within the CESAR project:

- Development of the corpus (incl. annotation, verification, structuring)
- Clearance of license terms

The corpus is distributed under META-SHARE NoRedistribution Non-Commercial license for free.

8.24. The Bulgarian-English Sentence- and Clause-Aligned Corpus

The Bulgarian-English Sentence- and Clause-Aligned Corpus (BulEnAC) is an excerpt from the Bulgarian-English Parallel Corpus – a part of the Bulgarian National Corpus (BulNC) of app. 280.8 million tokens and 8.2 million sentences for Bulgarian and 283.1 million tokens and 8.9 million sentences for English. The Bulgarian-English Parallel Corpus has been processed at several levels: tokenisation, sentence splitting, lemmatisation. The processing has been performed using the Bulgarian language processing chain for the Bulgarian part and Apache OpenNLP and Stanford CoreNLP for the English part.
The BulEnAC consists of 176,397 tokens for Bulgarian and 190,468 for English (366,865 tokens altogether). The BulEnAC comprises 14,667 Bulgarian sentences (12.02 words per sentence on average) and 15,718 English sentences (12.11 words per sentence). The average number of clauses in a sentence in the Bulgarian part is 1.67 compared to 1.85 clauses per sentence for the English part.
The texts are distributed over five broad categories, called 'styles': administrative, fiction, science, journalism, and subtitles. The corpus is represented in XML format and is supplied

with various linguistic annotation – monolingual for both Bulgarian and English (sentence splitting, tokenisation, lemmatisation, POS and grammatical tagging), and parallel (sentence and clause alignment).

The following work has been carried out for the resource within the CESAR project:

- Development of the corpus (building, annotation, verification, tests)
- Clearance of license terms

The corpus is distributed under META-SHARE NoRedistribution Non-Commercial license for free.

8.25. Mutilingual dictionaries

The set of multilingual dictionaries covers all pairs of languages among the following: Bulgarian, English, German, Romanian, Greek, and Polish. The main source of the dictionaries is Wikipedia – translations of article titles and category labels. The dictionaries include single words, MWEs and phrases but are predominantly phrase-to-phrase.
The following sets of dictionaries are included in the pack:

- General bilingual dictionaries for each pair of languages.
- Bilingual dictionaries of personal names for each pair of languages.
- Bilingual dictionaries of organisations for each pair of languages
- Bilingual dictionaries of toponyms for each pair of languages.

The dictionaries are stored in plain text format for easy and flexible storage and processing.
The following work has been carried out for the resource within the CESAR project:

- Development of the dictionaries
- Clearance of license terms

The dictionary is distributed under META-SHARE NoRedistribution Non-Commercial license for free.

8.26. Bulgarian MWE dictionary

The Bulgarian dictionary of MWEs includes 27,744 MWEs altogether which are divided into 13 categories based on their idyomaticity which is evaluated with respect to the following features:

- whether the MWE is a named entity;
- whether the MWE contains a reference to a named entity;
- the degree to which the meaning of the MWE is compositional and transparent.

The MWEs are extracted from several sources: Wikipedia, the Thesaurus of Bulgarian (1994) and other printed dictionaries and electronic corpora. The MWEs are manually verified and classified into categories.

The following work has been carried out for the resource within the CESAR project:

- Development of the dictionary up to its present size
- Verification
- Clearance of license terms

The dictionary is distributed under META-SHARE NoRedistribution Non-Commercial license for free.

8.27. bgMWE – tool for MWE recognition

bgMWE is a tool for corpus processing and MWE recognition and tagging. It is developed in Java and is platform independent. bgMWE comprises a set of modules which can be applied for particular NLP tasks. It is largely language independent and can work either in resource-light mode, or its performance can be boosted by employing lexical resources. The system includes the following modules:

- Web crawler for Wikipedia.
- Extraction of lexical data – lists of words and MWEs.
- Converter between formats – vertical format, XML, etc.
- Preprocessing module – applying a chunker, a tagger, etc.
- Collection of frequency data.
- MWE recognition and tagging.

The following work has been carried out for the resource within the CESAR project:

- Development of the dictionary
- Verification
- Testing methods for efficiency improvement
- Clearance of license terms

The dictionary is distributed under META-SHARE NoRedistribution Non-Commercial license for free.

8.28. TextMatch

TextMatch is a web service that computes similarity between two documents using powerful linguistic tools. The service gives different measures of similarity while returning an overall percentage for the probability of similarity between two documents. TextMatch compares documents in different formats (such as MS Office, OpenDocument, Portable Document, Electronic Publication, HyperText Markup Language, Rich Text Format, different text formats). TextMatch recognizes the language of the document using two-stage language detection system. Tokenizers, lemmatizers and other analyzers are utilized for English, Bulgarian, German, French, and Russian. A language independent comparison algorithm is used if one of the uploaded documents is in another language.

The following work has been carried out for the resource within the CESAR project:

- Development of the web service
- System deployment
- Tests and verification
- Development of a webpage
- Clearance of license terms

The results are offered under META-SHARE NoRedistribution Non-Commercial license for free.

8.29. Bulgarian grammar checker web service

The Bulgarian Grammar checker is based on a language model derived from the frequency list of the annotated Bulgarian National Corpus. It checks 893,626,788 3-grams with POS tags, including punctuation. The results show the probability of an arbitrary 3-gram with part-of-speech tags to be valid in the language model. The language model is executed in the form of finite automata. For each sentence, the model consecutively applies 3-grams, and those that are below the threshold are flagged as potential errors.
The following work has been carried out for the resource within the CESAR project:

- Further development of the web service.
- Deployment of the system
- Tests and verification
- Clearance of license terms

The results are offered under META-SHARE NoRedistribution Non-Commercial license for free.

8.30. N-grams from Bulgarian National Corpus

BgNgrams lists are extracted from the current version of the Bulgarian National Corpus (with a core Bulgarian part containing over 1.2 billion words). The n-grams involves both lemmas (n-gram lemma) and word forms (n-gram word form). n-grams can be 1-grams, 2-grams, 3-grams, 4-grams, 5-grams. The n-gram language models (1-5) are in the standard ARPA text and binary format.
The following work has been carried out for the resource within the CESAR project:

- Extraction of the n-grams
- Selection
- Verification
- Clearance of license terms

The results are offered under META-SHARE NoRedistribution Non-Commercial license for free.

8.31. Bulgarian Automatic Collocation Dictionary

The Automatic Collocations Dictionary is produced by Lexical Computing Limited using a corpus of the language, lemmatised and pos-tagged. The corpus is load into the Sketch Engine (http://www.sketchengine.co.uk). A 'sketch grammar' (of regular expressions over part-of-speech tags), when applied to a corpus, identifies a set of collocations, e.g. <headword, grammatical relation, collocate> triples. For all lexical words of sufficient frequency, all collocations they participate in are listed.
The following work has been carried out for the resource within the CESAR project:

- A web component of the Bulgarian National Corpus comprising 419 million words, as prepared, lemmatised and tagged by the Institute for the Bulgarian Language (IBL) was loaded into Sketch Engine.
- The sketch grammar prepared by IBL was applied.
- The resulting dictionary has entries for 31,011 headwords, with an average of 4.2 collocations per headword.
- The entry for each collocation was designed to include pointers to its corpus examples on the Sketch Engine website.
- The results were verified.
- The license terms were resolved.

The results are offered under META-SHARE NoRedistribution Non-Commercial license for free.

8.32. Bibliography of Bulgarian Lexicology, Phraseology and Lexicography

The database of Bibliography of Bulgarian Lexicology, Phraseology and Lexicography contains bibliographic units for published from 1950 onwards studies of Bulgarian lexis, phraseology and lexicography. The database contains about 6,600 bibliographic records (over 200 monographs, 29 collections, 32 textbooks, 1,968 articles in different collections, 3,007 articles in periodicals, 186 dissertations, 202 reviews of scientific papers, 261 reviews of dictionaries, etc.). The database includes publications in Bulgarian, Russian, Ukrainian, Belarusian, Polish, Czech, Slovak, Serbian, Romanian, German, French, English, Italian, Spanish, and Portuguese. Each bibliographic unit is accompanied by keywords (from a list of about 250 keywords) focused on the content of the publication.
The following work has been carried out for the resource within the CESAR project:

- Excerption of bibliographic records of paper and electronic publications.

- Attribution of keywords reflecting the content of the publication (using a list of 250 keywords).
- Verification and editing of the data.
- Entering the data into a database.

The results are offered under META-SHARE NoRedistribution Non-Commercial license for free.

9. LSIL resources

9.3. Slovak Morphology Database

The following work has been carried out for the resource within CESAR since the inclusion of the corpus in the 2nd batch of resources:

- several thousand entries from the new Dictionary of Contemporary Slovak have been added, up to the current size of 96 thousand lemmas.
- new version has been released via META-SHARE

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

9.4. Slovak-English Parallel Corpus (all)

The corpus consists of parallel Slovak and English texts, with automatic lemmatization, morphological analysis (for Slovak), POS tagging (for English). The corpus consists of two parts – the subcorpus of "fiction" and the free subcorpus. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.

Since the 1st batch of resources, a new version of the corpus has been created, substantially increasing the amount of fiction texts in the corpus (from 1.5 million sentence pairs to 4 million) and the "free" part of less copyright-encumbered texts has been added to the corpus, resulting in 10 million sentence pairs in total.

9.5. Slovak-Czech Parallel Corpus (all)

Parallel Slovak-Czech corpus is a corpus of sentence aligned texts. Corpus consists of two parts: the subcorpus of fiction and the free subcorpus. The Slovak texts are morphologically annotated and disambiguated using the system applied in the Slovak National Corpus, Czech texts are annotated with the morče tagger. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.

Since the 1st batch of resources, a new version of the corpus has been created, adding the "free" part of less copyright-encumbered texts to the corpus, resulting in 6.4 million sentence pairs in total.

9.14. Slovak-Czech Parallel Corpus (free)

Parallel Slovak-Czech corpus is a corpus of sentence aligned texts of freely downloadable texts. The Slovak texts are morphologically annotated and disambiguated using the system applied in the Slovak National Corpus, Czech texts are annotated with the morče tagger.

This is a new resource, texts of which are also included in the *Slovak-Czech Parallel Corpus (all)*. It has been released separately, because the license of these texts allows users to download annotated and aligned texts. The size of the corpus is 5.7 million sentence pairs.

9.15. Slovak-English Parallel Corpus (free)

The corpus consists of parallel Slovak and English freely downloadable texts, with automatic lemmatization, morphological analysis (for Slovak), POS tagging (for English). The corpus consists of original English language books and their Slovak translations.

This is a new resource, texts of which are also included in the *Slovak-English Parallel Corpus (all)*. It has been released separately, because the license of these texts allows users to download annotated and aligned texts. The size of the corpus is 6 million sentence pairs.

9.16. Slovak Terminology Database

Slovak Terminology Database is a database of 6 000 terms from 23 fields. The database has been extended by several hundred terms, and a new field (computer science) has been added to it.

9.17. Corpus of Informational Texts prim-6.0-inf

INF is a corpus consisting of informational text (mostly newspapers and journals). This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed. This is part of the new release of Slovak National Corpus and its subcorpora. Since the previous version, the size of the corpus increased from 515 to 889 million tokens, and the corpus has been released to META-SHARE as a separate resource.

9.18. Corpus of Professional Texts prim-6.0-prf

PRF is a corpus consisting of professional text (including popular science). The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed. This is part of the new release of Slovak National Corpus and its subcorpora. Since the previous version, the size of the corpus increased from 82 to 106 million tokens, and the corpus has been released to META-SHARE as a separate resource.

9.19. Corpus of Fiction prim-6.0-img

MG is a corpus of fiction. The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed. This is part of the new release of Slovak National Corpus and its subcorpora. Since the previous version, the size of the corpus increased from 99 to 114 million tokens, and the corpus has been released to META-SHARE as a separate resource.

9.20. Corpus of Original Slovak Texts prim-6.0.sk

This is a corpus of original Slovak texts (no translations). The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed. This is part of the new release of Slovak National Corpus and its subcorpora. Since the previous version, the size of the corpus increased from 508 to 905 million tokens, and the corpus has been released to META-SHARE as a separate resource.

9.21. Corpus of Original Slovak Fiction prim-6.0-skimg

This is a corpus of original Slovak fiction (no translations). The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed. This is part of the new release of Slovak National Corpus and its subcorpora. Since the previous version, the size of the corpus increased from 32 to 35 million tokens, and the corpus has been released to META-SHARE as a separate resource.

9.22. Corpus of Slovak Texts from the Years 1955 to 1989

R55AZ89 is a corpus containing texts from the years 1955 to 1989. The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed. This is part of the new release of Slovak National Corpus and its subcorpora. This is a new resource which has not been available previously, and the corpus has been released to META-SHARE as a separate resource. The corpus contains 63 million tokens.

9.23. Corpus of Historical Slovak

Corpus of Historical Slovak contains texts from the 16th, 17th and 18th centuries. The corpus is a database of electronically processed texts published in Pramene k dejinám slovenčiny, I. - III; (Sources for the History of Slovak). This is a new resource, the size of the corpus is 371 thousand tokens.

9.24. Lithuanian WordNet

Lithuanian WordNet is a lexical database including information about semantic relations of Lithuanian words. It is aligned with the Princeton 3.0 WordNet. This is a new resource, the size of the WordNet is 10 thousand synsets.

9.25. Dictionary of Slovak Collocations. Nouns

Dictionary of noun collocations. It is the only one existing collocation dictionary in Slovakia. It is a dictionary of not only phrasemes, but also of common word collocations. This is an external resource created at the Faculty of Philosophy of St. Cyril and Methodius University in Trnava. So far it is available through web interface and under restrictive licensing, but the

authors promised to release the dictionary under Open Content license after the publication of the dictionary in book format.

9. 26. Dictionary of Slovak Collocations. Adjectives

Dictionary of adjective collocations. It is a dictionary of not only phrasemes, but also of common word collocations. This is an external resource created at the Faculty of Philosophy of St. Cyril and Methodius University in Trnava. So far it is available through web interface and under restrictive licensing, but the authors promised to release the dictionary under Open Content license after the publication of the dictionary in book format.

9.27. Slovak WordNet

Slovak WordNet is a lexical database including information about semantic relations of Slovak words. It is aligned with the Princeton 3.0 WordNet. This is a new resource, not available previously. The size of the WordNet is 20 thousand synsets.

9.28. Slovak National Corpus prim-6.0

The Slovak National Corpus is a representative corpus of contemporary Slovak language written texts published since 1955 (1953 being the time of most recent substantial Slovak language orthography reform). The corpus is automatically lemmatised and MSD tagged. The documents are annotated with their genre, style and other bibliographic information. There are specialised subcorpora containing fiction, informational texts, professional texts, original Slovak fiction, texts written from 1955 to 1989, and a balanced subcorpus. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed. This is part of the new release of Slovak National Corpus and its subcorpora. Since the previous version, the size of the corpus increased from 719 to 1155 million tokens.

9.29. Balanced Slovak Corpus prim-6.0-vyv

VYV is a balanced corpus with respect to text type. It contains 1/3 fiction, 1/3 informational text, 1/3 professional text (including popular science). The texts were selected from the Slovak National Corpus according to their style-genre annotation. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed. This is part of the new release of Slovak National Corpus and its subcorpora. Since the previous version, the size of the corpus increased from 247 to 313 million tokens, and the corpus has been released to META-SHARE as a separate resource.

The previous version of the corpus has remained available as well, because of the users who started long term projects and depend on stable version of reference corpora.

9.30. Language model prim-6.0-sane

This is a language model from the Slovak National Corpus. This lowercased model is from a

1200 million token collection. Language model is in the iARPA format, using witten-bell smoothing, created by the IRSTLM Tooklit. The previous version of the model has remained available as well, because of the users who started long term projects and depend on stable version of reference corpora.

9.31. Language model prim-6.0-inf

This is a lowercased language model of journalistic style. The model is built on corpus of 889 million tokens. The language model is in iARPA format, using witten-bell smoothing, and was created by the IRSTLM Tooklit. The previous version of the model has remained available as well, because of the users who started long term project and depend on stable version of reference corpora.

9.32. Language model prim-6.0-vyv

This is a lowercased language model of balanced language. The model is built on the balanced Slovak corpus of 313 million tokens. The language model is in iARPA format, using witten-bell smoothing, and was created by the IRSTLM Tooklit. The previous version of the model has remained available as well, because of the users who started long term projects and depend on stable version of reference corpora.

9.33. Parallelum Slovaco-Latinum Corpus

Parallel Slovak-Latin Corpus is a database of Latin texts translated into Slovak. This is a new resource, the size of the corpus is 22 thousand Latin and 26 thousand Slovak sentences.

9.34. n-grams from Slovak National Corpus

Set of n-grams extracted from the Slovak National Corpus for $1 \leq n \leq 4$. The resource contains all unique n-grams preceded and sorted by number of occurrences. There are separate files for case sensitive and for lowercased tokens. Previously, the n-gram files have been limited to trigrams, this release increased the order to the 4-grams, and is built on the new version of the Slovak National Corpus (6.0).

9.35. Multilingual Glossary of Synsets

Multilingual glossary of synsets is a common resource produced by several CESAR partners – HASRIL, FFZG, UBG, IBL, LSIL.It has been created by mapping several existing WordNets to the Princeton WordNet v. 3.0. and contains synsets (nouns and adjectives) in Bulgarian, Croatian, Hungarian, Serbian and Slovak, together with the links to the English WordNet.

The creation of this resource is a direct application of the cross-lingual alignment described in *Actions – Aligning resources across languages*. Existing WordNet databases available for partners' respective languages has been aligned with Princeton WordNet v3.0 (in some cases

– Hungarian – a new alignment had to be generated) and a glossary of nouns and adjectives has been generated. The resource contains 2296 entries and has been released under permissive Princeton WordNet license.

9.36. Automatic Collocation Dictionary of Slovak

The Automatic Collocation Dictionary of Slovak has been produced by Lexical Computing Ltd., based on the web corpus of Slovak language and Slovak word sketches (developed at LSIL). The entry for each collocation includes pointers to its corpus examples on the Sketch Engine website.

The entry contains major grammatical relations that the headword occurs in and collocates it occurs within that grammatical relation.

The following work has been carried out for the resource within CESAR:

- A distribution package of the corpus prepared by Lexical Computing.
- Corpus metadata description created to maintain META-SHARE compliance.
- The dictionary has been publicly released under Open Database License v1.0.