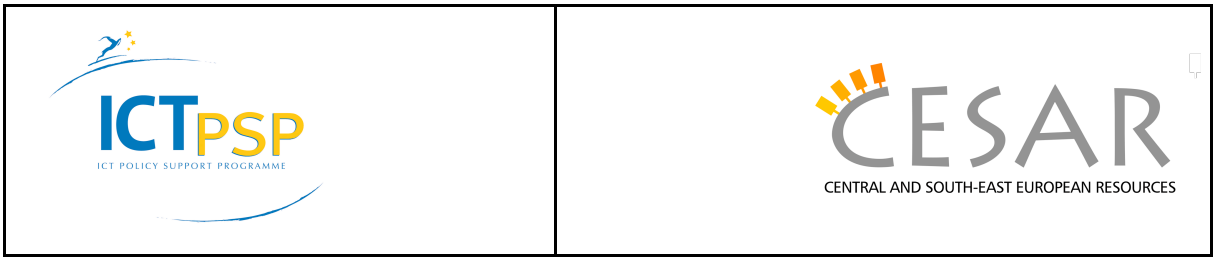

CESAR

Central and South-East European Resources

Project no. 271022

Deliverable D3.2.-B

Second batch of language resources: actions on resources

Version No. 1.2

31/07/2012

Document Information

Deliverable number:	D3.2.-B
Deliverable title:	Second batch of language resources: actions on resources
Due date of deliverable:	31/07/2012
Actual submission date of deliverable:	31/07/2012
Main Author(s):	György Szaszák (BME-TMIT)
Participants:	Cvetana Krstev, Duško Vitas (UBG) Maciej Ogrodniczuk, Adam Przepiórkowski, Łukasz Degórski, Katarzyna Głowińska, Michał Lenart, Leszek Manicki, Marcin Miłkowski, Agata Savary, Filip Skwarski, Joanna Świetlicka, Zygmunt Vetulani, Adam Wardyński, Dawid Weiss, Marcin Woliński, Bartosz Zaborowski (IPIPAN) Radovan Garabík, Adriána Žáková (LSIL) Mátyás Bartalis, Gábor Olaszy, András Balog (BME-TMIT) Tibor Pintér, Veronika Vincze, Dániel Varga (HASRIL) Svetla Koeva, Tsvetana Dimitrova (IBL) Marko Tadić (FFZG)
Internal reviewer:	Tamás Váradi (HASRIL)
Workpackage:	WP3
Workpackage title:	Enhancing language resources
Workpackage leader:	IPIPAN
Dissemination Level:	PP
Version:	1.2
Keywords:	Upload, interoperability, standardization, harmonization, upgrade, extension, linking

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
1.2	30/07/2012	Proofreading	Tamás Váradi Katalin Tóth (HASRIL)	Tibor Pintér (HASRIL)	proofreading
1.1	28/07/2012	Pre-final	György Szaszák	All partners	completion
1.0	03/07/2012	draft	György Szaszák (BME-TMIT)	From all Partners	core material

Executive summary
This deliverable entitled D3.2.-B. is a supplement for the publicly available deliverable D3.2., treating second 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 second upload batch. Work for resources or tools uploaded in the first batch, but extended or updated for the second batch is also documented. Paragraph numbering corresponds to resource Ids (in metadata description), hence paragraph numbering is not continuous.

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 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 and open as possible, depending 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 than in the first one are related to extension, linking of resources. For resources released in the first batch, the emphasis was put, not exclusively, 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 first batch resources, 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 was based on those made available within task 3.1 to further enhance a smaller, but well-defined set of resources. The following rationale was applied:

- the extension of resources should provide considerable value to the community, at least on the 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 type of tool 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 it was found that a certain less-developed, but still, at least within one language community very popular tool can benefit from the enhancement based on their well-developed equivalent, its enhancement was also considered (provided that no extensive work will be required and that the latter tool cannot be used as a building block in further applications of the former tool,
- 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 are to be found in deliverables related to WP2.

0.3. Actions – Aligning resources across languages

Cross-lingual alignment of resources, as the most demanding task, will 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 the scope of interest of the CESAR consortium 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 2nd batch

The 2nd upload batch and related activities were grouped in CESAR as follows:

- set up META-SHARE node version 2.1, provide metadata schemes version 2.1. 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 the resource point of view) – presented and listed in D3.2.
- actions on resources as required in WP3 (upgrade, extension, standardization, harmonization, etc.) – presented here in D3.2.-B.

This document is an extension to deliverable D3.2. and its paragraph numbering follows that of D3.2. This means that related activities to a given LRT described under a given section in D3.2. 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 2. 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 the CESAR consortium have not only fulfilled but even exceeded their obligations as some resources scheduled for later batches have already been included in batch 1 and now in batch 2. Partners who have added extra LRTs to batch 2 are planning to either carry on further actions on the already released LRTs – which may significantly increase their usability (be on a higher level of technology, wider range of usability, etc.) – or spend their saved time/costs on exploring new resources or tools to be included into the META-SHARE network.

0.5 Key Performance Indicators

The mid-term project review suggested assessing the performance and effort of consortium partners with pre-defined Key Performance Indicators (KPI). Some of these are closely linked to the resources and are presented in subsequent tables for batch 2.

0.5.1 Resource statistics

KPI – Resources I.	Value
1) Total number of items	108
2a) of which % from outside the consortium	21.5
2b) of which % not available from ELRA and/or LDC	69.2
2c) of which % cleared with owner/holder and posted on the Net (Share and the like)	73.8

Table 2. KPI for resources in second batch

0.5.2 IPR statistics

KPI – IPR	Value
1) Total number of licences concluded	84
2) of which % are new i.e. no previous licence	60.0
3) of which % are for free research use/reuse	63.10

Table 3. KPI for IPR related activity in second batch

0.5.3 Distribution of resources

	Corpus	Lexicon/Conceptual	Language desc.	Technology/tool	All types
Text	41	26	1	24	92
Audio	14	1	0	0	15
Video	0	1	0	0	1
Image	0	0	0	0	0
All media	55	28	1	24	108

Table 4. KPI – overall distribution of resources involved in batch 2

0.5.4 Language coverage

Bulgarian

	Corpus	Lexicon/Conceptual	Language desc.	Technology/tool	All types
Text	2	3	0	5	10
Audio	0	0	0	0	0
Video	0	0	0	0	0
Image	0	0	0	0	0
All media	2	3	0	5	10
Effort spent on documentation and annotation improvement (% of total WP3 PM)	10	2	0	8	20

Table 5.1-BU KPI – Distribution (number) of resources involved in batch 2 for Bulgarian

	Corpus	Lexicon/Conceptual	Language desc.	Technology/tool
Text	90	80	0	90
Audio	0	0	0	0
Video	0	0	0	0
Image	0	0	0	0

Table 5.2-BU KPI – Language coverage (%) in batch 2 for Bulgarian

Croatian

	Corpus	Lexicon/Conceptual	Language desc.	Technology/tool	All types
Text	5	0	0	2	7
Audio	0	0	0	0	0
Video	0	0	0	0	0
Image	0	0	0	0	0
All media	0	0	0	0	0
Effort spent on documentation and annotation improvement (% of total WP3 PM)	5	0	0	2	7

Table 5.1-CR KPI – Distribution (number) of resources involved in batch 2 for Croatian

	Corpus	Lexicon/Conceptual	Language desc.	Technology/tool
Text	90	0	0	10
Audio	0	0	0	0
Video	0	0	0	0
Image	0	0	0	0

Table 5.2-CR KPI – Language coverage (%) in batch 2 for Croatian

Hungarian

	Corpus	Lexicon/Conceptual	Language desc.	Technology/tool	All types
Text	8	2	0	3	13
Audio	9	0	0	0	9
Video	0	0	0	0	0
Image	0	0	0	0	0
All media	17	2	0	3	22
Effort spent on documentation and annotation improvement (% of total WP3 PM)	20	0	0	0	20

Table 5.1-HU KPI – Distribution (number) of resources involved in batch 2 for Hungarian

	Corpus	Lexicon/Conceptual	Language desc.	Technology/tool
Text	40	40	0	90
Audio	70	25	0	0
Video	10	0	0	0
Image	0	0	0	0

Table 5.2-HU KPI – Language coverage (%) in batch 2 for Hungarian

Polish

	Corpus	Lexicon/Conceptual	Language desc.	Technology/tool	All types
Text	10	11	1	9	31
Audio	2	0	0	0	2
Video	0	0	0	0	0
Image	0	0	0	0	0
All media	12	11	1	9	33
Effort spent on documentation and annotation improvement (% of total WP3 PM)	40	20	1	30	71

Table 5.1-PO KPI – Distribution (number) of resources involved in batch 2 for Polish

	Corpus	Lexicon/Conceptual	Language desc.	Technology/tool
Text	80	75	20	80
Audio	60	0	30	0
Video	0	0	0	0
Image	0	0	0	0

Table 5.2-PO KPI – Language coverage (%) in batch 2 for Polish

Serbian

	Corpus	Lexicon/Conceptual	Language desc.	Technology/tool	All types
Text	6	2	0	2	10
Audio	0	0	0	0	0
Video	0	0	0	0	0
Image	0	0	0	0	0
All media	6	2	0	2	10
Effort spent on documentation and annotation improvement (% of total WP3 PM)	15	5	0	5	25

Table 5.1-SR KPI – Distribution (number) of resources involved in batch 2 for Serbian

	Corpus	Lexicon/Conceptual	Language desc.	Technology/tool
Text	50	25	0	28.5
Audio	0	0	0	0
Video	0	0	0	0
Image	0	0	0	0

Table 5.2-SR KPI – Language coverage (%) in batch 2 for Serbian

Slovak

	Corpus	Lexicon/Conceptual	Language desc.	Technology/tool	All types
Text	10	8	0	3	21
Audio	3	1	0	0	4
Video	0	1	0	0	1
Image	0	0	0	0	0
All media	13	9	0	3	25
Effort spent on documentation and annotation improvement (% of total WP3 PM)	20	5	0	0	25

Table 5.1-SL KPI – Distribution (number) of resources involved in batch 2 for Slovak*

	Corpus	Lexicon/Conceptual	Language desc.	Technology/tool
Text	70	22	0	100
Audio	66	0	0	0
Video	0	0	0	0
Image	0	0	0	0

Table 5.2-SL KPI – Language coverage (%) in batch 2 for Slovak

* Slovak database of gestures is included in both audio and video (by LSIL)

1. HASRIL resources

1.11. SzegedParalell

The English-Hungarian parallel corpus contains texts that are manually aligned at the paragraph and sentence levels.

The actions carried out within the CESAR project cover the followings:

- The website of the corpus has been updated
- Alignment errors have been corrected
- The documentation of the resource has been improved

Updating the website will result in an easier distribution process of the corpus and the correction of alignment errors will also yield a more accurate database.

1.12. SzegedParalellFX

The SzegedParalellFX corpus is an English-Hungarian parallel corpus, in which light verb constructions are manually annotated.

The actions carried out within the CESAR project cover the followings:

- The website of the corpus has been updated
- Annotation errors have been corrected
- The documentation of the resource has been updated
- The documentation has been translated to English
- Basic statistical data have been prepared on the frequency of light verb constructions
- The documentation of the resource has been updated
- A list of light verb constructions has been made available for both languages

Updating the website and the documentation will contribute to the easier distribution of the resource and their English translation will enhance the international usage of the data. Error correction was necessary for having more reliable data and the consistency of the annotation can be also improved in this way. Basic statistical data will help quantitative research on light verb constructions and the lists will enhance various NLP applications.

1.13. Szeged Treebank FX

The Szeged Treebank FX is a database in which words are morphosyntactically tagged, sentences are syntactically parsed and light verb constructions are also manually annotated.

The actions carried out within the CESAR project cover the followings:

- The website of the corpus has been updated
- Annotation errors have been corrected
- The annotation was mapped to the dependency version of the treebank
- Basic statistical data have been prepared on the frequency of light verb constructions

- The documentation of the resource has been updated
- The documentation has been translated to English
- A list of light verb constructions has been made available

Updating the website and the documentation will contribute to the easier distribution of the resource and their English translation will enhance the international usage of the data. Error correction was necessary for having more reliable data and the consistency of the annotation can be also improved in this way. Basic statistical data will help quantitative research on light verb constructions and the list will enhance various NLP applications.

1.14. Hungarian WSD Corpus

The Hungarian WSD corpus contains 300-500 occurrences of 39 word forms that are word sense disambiguated.

The actions carried out within the CESAR project cover the followings:

- The website of the corpus has been updated
- A few XML errors have been corrected
- The documentation of the resource has been improved

Updating the website will result in an easier distribution process of the corpus and the correction of XML errors will also yield a more accurate database.

1.15. Szeged Criminal NE Corpus

The corpus contains texts on criminal offences which are annotated for named entities.

The actions carried out within the CESAR project cover the followings:

- The website of the corpus has been updated
- A few annotation errors have been corrected

Updating the website will result in an easier distribution process of the corpus and the correction of annotation errors will also yield a more accurate database.

1.16. Hungarian Verb Phrase Constructions

Hungarian Verb Phrase Constructions is a list of verb phrase constructions (VPC) automatically extracted from the Hungarian National Corpus.

The actions carried out within the CESAR project cover the followings:

- Conversion of the original database to standard represetation using LMF (Lexical Markup Framework)

1.17. Hungarian NER Corpus based on Wikipedia

The text of the corpus is automatically generated from Hungarian Wikipedia articles which contains Named Entity (NE) tagging according to the CoNLL standard (Person, Organization, Location and Miscellaneous), and additional morphological annotation. The corpus is the largest ever NE tagged corpus for Hungarian, which can be used for training and testing NE recognizer applications.

The actions carried out within the CESAR project cover the followings:

- technology related upgrade (download, parsing and cleaning of the XML-files, NE-labeling)
- enhancement of the NE-tagger
- enhancement of the documentation

1.18. Hungarian Opinion-Tagged Sentence Bank

The OpinHuBank is a human-annotated resource for researching, evaluating and developing opinion mining systems for Hungarian. The resource consists of several thousand sentences selected from Hungarian online newswire, blogs and social media.

The actions carried out within the CESAR project cover the followings:

- upgrade of the NER tools
- annotation of the polarity of each sentence
- enhancement of the documentation

1.19. Hungarian Language Processing Tools in NooJ

The following upgrades and extensions were made on Hungarian tools in NooJ:

Upgrade of the main dictionaries

Nominal dictionaries (noun.dic and noun.flx files):

- Derivational suffixes were added to the nominal paradigm (eg. *-talan, -mentes*: noun to adjective *-zik*: noun to verb derivational suffixes were added).
- Further inflected forms of the derived forms were generated.
- Complex inflected forms were also generated. The complete list contains over 96 million word forms!
- A large sample (over 120.000 word forms) from the generated word forms were hand validated. While the validation process bugfixing and error categorisation was made. Some of them were resulted by the generation process - these errors were corrected in the *noun.flx* file. Some errors were resulted by the morphological encoding.

Verbal dictionaries (verb.dic, verb.flx, verb.nod)

- Derivational suffixes were added to the verbal dictionaries.

- A large sample (over 30.000 word forms) from the generated word forms were hand validated. The errors were categorised. Some of them were the result of the generation process these errors were corrected in the verb.flx file. Some errors were the result of the morphological encoding.

The words which should not have affixes were added into noaffix.dic file.
New dictionaries were compiled by using the .dic and .flx files.
Complex inflected forms were generated.

1.20. hunner tool

hunner is a tool constructed for performing any kind of supervised sequential sentence tagging tasks. It has been used for NP chunking, Named Entity Recognition, and clause chunking.

The following actions were made on hunner:

- Restructuring the package.
- Evaluations to find the most accurate maximum entropy models to be provided together with the tool (for Hungarian and English NP-chunking, general chunking, named entity recognition and clause detection).
- Documentation translated from Hungarian and extended.

1.21. hunpars tool

hunpars is a syntactic analyzer developed for Hungarian language. hunpars can explore the syntactic structure of the simple Hungarian sentences. Works on the hunpars tool covered

- metadata collection
- minor bugfixes
- refreshing of the English documentation
- documentation enhancement

1.22. hunpos tool

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

- metadata collection
- minor bugfixes
- refreshing of the English documentation
- documentation enhancement

2. BME-TMIT resources

2.7. Hungarian MTBA

Hungarian MTBA is issued from a project for the creation of the fixed line and mobile telephone voices based Hungarian speech database.

The goal of the project was collecting speech telephone database, in which some major dialectal variants are represented. This database provided a realistic base both for the training and testing of the present-day teleservices, and - because of the phonetically richness - the training of real speaker independent speech recognizers. The database contains records based on the definition in SpeechDatE for the dialectical, age and sex balance and vocabulary. Important and different from the SpeechDatE database is, that the phonetically rich sentences and words have been segmented and labelled at phoneme level. Thus the database allows one to train phoneme based recognizers. In planning the corpus, we took into consideration not only the variety of the dialectical aspects but the special characteristics of the Hungarian language too. Since the Hungarian is an agglutinative language, we needed to create a larger vocabulary in some categories, than it was mandatory. We tried to pay extra attention to the topic 'phonetically rich sentences and words', to create a phonetically well balanced speech database for text independent speech recognizers. A detailed statistical analysis was prepared to examine the statistics of phonemes, diphones, triphones and syllables.

Actions carried out within CESAR for batch 2:

- Documentation improvement
- Metadata annotation
- IPR clearance
- Phoneme level segmentation extension and checking of segmentations

2.8. Hungarian MRBA

The Hungarian Reference Speech Database (MRBA) was developed at the Laboratory of Speech Acoustics of of the Budapest University of Technology and Economics (BME) in collaboration with the Institute of Informatics of the University of Szeged. The main goal was to develop a speech database that contains continuous read speech, so that the database can be used for training and testing of PC-based automatic speech recognisers. During the planning of the corpus, we took into consideration the special characteristics of the Hungarian language. Since Hungarian is an agglutinative language, we needed to create a larger vocabulary in some categories than it is mandatory. We tried to pay an extra attention to the topic 'phonetically rich sentences and words', to create a phonetically well balanced speech database for text independent speech recognizers. A detailed statistical analysis was prepared, to examine the statistics of phonemes, diphones, triphones and syllables.
In this way, every speaker had to read 12 different sentences and 12 different words that had

no connection with the sentences. The database contains utterances read by 332 different speakers.

The utterances were recorded in acoustically different locations, such as office, laboratories, home. The database contains utterances recorded simultaneously with two different systems. One of these systems was considered the reference system. This reference system contained a laptop, an external sound card and a good quality condenser microphone. The reference system was being unchanged until the database hadn't been finished. In case of the other system, we changed the microphones, sound cards, PC-s.

To cover the dialects spoken in Hungary, we made records in four different locations of the country. The database is balanced according to gender, age and dialects.

Every spoken utterance has been labeled, thus every wave (16kHz, 16bit, mono) file has got a label file, which contains informations about the parameters of the record and the ortographical transcription of the spoken material. Almost one third of the database (100 speakers' utterances) was manually segmented and labelled at phoneme level, using SAMPA codes.

Actions carried out within CESAR for batch 2:

- Documentation improvement
- Metadata annotation
- IPR clearence
- Phoneme level segmentation extension and checking of segmentations

2.9. Hungarian Telephone Speech Call Center Database

The Hungarian Phone Speech Call Center Database is a telephone speech database containing discourses between the operators of a service provider company and its clients. Ortrhographic transcription is provided. Emotions are also labelled. A derivative of this database called Hungarian Speech Emotion Database is also available from META-SHARE resulting from batch 1, with free academic use.

Actions carried out within CESAR for batch 2:

- Database anonymisation: removal of names, addresses and other personal data from the audio signal
- Audio and text encoding standardization, upgrade for interoperability
- Database structure cleanup (creation of a directory structure)
- Splitting utterances into individual speech files
- Generating uniform transcriptions
- Generating and enhancing documentation
- Clearance of license terms
- Metadata description

Hungarian Telepphone Speech Call Center Database was a raw set of partly emotional speech recordings collected via telephone in a call center. It contained much personal data (names, addresses, telephone numbers, e-mail addresses, passwords etc), hence anonymisation prior to sharing was inevitable. Anonymisation was carried out by a self-made script followed by hand-made proofreading.

Audio encoding was either A-law 8 kHz 8 bit or Linear 8 kHz 16 bit or Linear 8 kHz 12 bit. This was set A-law 8 kHz 8 bit for all utterances. Text encoding of emotion labels varied UTF-8, UTF-16 and Latin-2. This was set UTF-8 uniformly. Some utterances of poor quality have been dropped.

A database structure was created as follows: each folder contains a conversation between the operator(s) and the client(s). Utterances are split into individual files to allow for drop out low quality, erroneous or uninformative (silent, strong noise etc.) parts.

Emotion transcription labels were uniformized for batch 1 (they were previously transcribed using two (although not disjunct) sets). Each conversation has one associated transcription file covering all utterances within the directory.

For batch 2, transcriptions were generated, anonimized, and structured. Alignment to the speech signal has been also provided.

Poor documentation was improved and translated to English (from Hungarian).

2.10. Hungarian BABEL phonetic and prosodic segmentation and syntactic analysis

BABEL database is an EUROM1 compatible database containing records from 60 speakers distributed in 3 speakers set (many, few, very few). The resource is available under META-SHARE via ELRA, this supplement is an add-on to the database. In order to use it, the BABEL database is necessary. Addons contain syntactic analysis, prosodic segmentation and ohoneme level segmentation of BABEL phonetically rich sentence utterances with corpus codes A, B or C (a total of 330 sentence utterances)..

Actions carried out within CESAR for batch 2:

- Creation of phoneme level segmentation (auto-alignment based on transcription)
- Syntatctic analysis (HunPars)
- Syntactic disambiguation and analysis cleanup and extension (manual)
- Creation of prosodic segmentation
- Structurization
- Documentation preparation
- Metadata labelling

2.11. Di-phone database for text-to-speech conversion in Hungarian

The Di-phone set (labelled wave form items) for Hungarian contains combinations of 38 sounds for TTS conversion. Besides, the Di-phone set can be used for educational purposes and in speech research.

Actions carried out within CESAR involved:

- Checking and optimizing the waveforms of the word items (silent part- word – silent part)
- 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
- Finalizing the sound sequence database from which the di-phone elements can be extracted.
- Filling and finalizing METADATA schemes and files

This database is currently offered under CLARIN RES, but license negotiations are ongoing and the resource is likely to be licensed later under META-SHARE NoRedistribution license.

2.12. Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection

This speech corpus consists of a phonetically balanced sentence set read by 10 speakers, with very high quality labelling.

Actions carried out within CESAR involved:

- Checking and optimizing the waveforms of the word items (silent part- word – silent part)
- 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
- Finalizing the sentence database.
- Filling and finalizing METADATA schemes and files

This database is currently offered under CLARIN RES, but license negotiations are ongoing and the resource is likely to be licensed later under META-SHARE NoRedistribution license.

2.13. Read Speech database in Hungarian

The read speech database contains sentences from weather forecast news. The sentence collection represents the four seasons. This database can be used for analysing speech characteristics in weather forecast news and also as the basic speech database of a corpus based Concept-to-Speech system.

Actions carried out within CESAR involved:

- Checking and optimizing the waveforms of the word items (silent part- word – silent part)
- 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
- Finalizing the sentence database.
- Filling and finalizing METADATA schemes and files

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

Database of portions of text and audio version of a Hungarian novel. The recordings are segmented between speech pauses, which do not necessarily correspond to sentence boundaries. The reading is mostly, but not completely accurate. Hence, an automatic speech recognizer was utilized to choose only those segments where there is a high match between the automatic recognition result and the original text. Thus the database comprises only those segments that are considered to have a reliable transcription. The database can be applied in speech technology research, phonetic, phonological research and for developing and testing speech recognition systems.

Actions carried out within CESAR for batch 2:

- Downloading text and audio data from Hungarian Electronic Library and librivox.org web site
- Converting the audio files
- Cutting the text and appending header sections to match the audio files
- Running the speech recognition engine. The audio data are segmented between speech pauses, and the segments are recognized.
- Comparing the segments' recognition result text with the manual transcription.
- Selecting the best matching segments
- Creating the metadata
- IPR clearence

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

Database of portions of text and audio version of a Hungarian piece of poetry. The recordings are segmented between speech pauses, which not necessarily correspond to sentence boundaries. The reading is mostly, but not completely accurate. Hence, an automatic speech recognizer was utilized to choose only those segments, where there is a high match between the automatic recognition result and the original text. Thus the database comprises only those segments that are considered to have a reliable transcription. The database can be applied in speech technology research, phonetic, phonological research and for developing and testing speech recognition systems.

Actions carried out within CESAR:

- Downloading text and audio data from Hungarian Electronic Library and librivox.org web site
- Converting the audio files
- Cutting the text and appending header sections to match the audio files

- Running the speech recognition engine. The audio data are segmented between speech pauses, and the segments are recognized.
- Comparing the segments' recognition result text with the manual transcription.
- Selecting the best matching segments
- Creating the metadata
- IPR clearence

2.16. Hungarian Parliamentary Speech and Aligned Text Selection Database

Database of recordings and official transcripts of Hungarian parliamentary speeches. The recordings are segmented between speech pauses, which not necessarily correspond to sentence boundaries. The official transcripts are not completely accurate, since the parliamentary transcribers correct most of grammatical mistakes and speech disfluencies. Hence, an automatic speech recognizer was utilized to choose only those segments, where there is a high match between the automatic and manual transcriptions. Thus the database comprises only those segments that are considered to have a reliable transcription. The database can be applied in speech technology research, phonetic, phonological research and for developing and testing speech and speaker recognition systems.

Actions carried out within CESAR for batch 2:

- Downloading text and video data from Hungarian Parliament web site
- Extracting audio data from the video files in the desired format
- Cleaning the text files to contain only speech data
- Running the speech recognition engine. The audio data are segmented between speech pauses, and the segments are recognized.
- Comparing the segments' recognition result text with the manual transcription.
- Selecting the best matching segments
- Creating the metadata
- IPR clearance

3. FFZG resources

3.11. Croatian Web Corpus (hrWaC)

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

- the boilerplate removal was done for the whole corpus with the WebContentExtractor
- the whole corpus has been lemmatised and MSD-tagged using CroTag system
- the corpus has been installed in NoSketchEngine which allows free on-line querying (http://faust.ffzg.hr/bonito2/run.cgi/first_form?corpname=hrwac)
- the corpus webpage has been produced (http://www.nljubesic.net/resources/corpora/hrwac/)
- documentation, compilation of metadata and corpus description

The Croatian Web Corpus (hrWaC) is the largest collected corpus for Croatian so far. It was collected in 2011-06 by crawling the whole .hr internet domain yielding ca 1.2 billion tokens. The corpus has been cleaned of HTML code, lemmatised and MSD-tagged automatically using CroTag system (Agić et al., 2008). . The compilation of the corpus is described in the TSD2011 paper Ljubešić, N., Erjavec, T. *hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene*. The morphosyntactically annotated and lemmatized corpus is distributed under the CC-BY-SA licence. It has been installed also in NoSketchEngine for free on-line querying: http://faust.ffzg.hr/bonito2/run.cgi/first_form?corpname=hrwac.

3.12. Slovene Web Corpus (slWaC)

The following work, similar to one done for hrWaC, has been carried out for the resource within CESAR:

- the boilerplate removal was done for the whole corpus with the WebContentExtractor
- the whole corpus has been lemmatised and MSD-tagged using ToTaLe system by Tomaž Erjavec
- the corpus has been installed in NoSketchEngine which allows free on-line querying (http://faust.ffzg.hr/bonito2/run.cgi/first_form?corpname=slwac)
- the corpus webpage has been produced (http://www.nljubesic.net/resources/corpora/slwac/)
- documentation, compilation of metadata and corpus description

Slovene Web Corpus (slWaC) is the the first version of the Slovene web corpus. It was collected by crawling the whole .si internet domain in 2011-06 yielding ca 380 million tokens. The corpus has been lemmatised and MSD-tagged automatically using ToTaLe system (Erjavec et al. 2005). The compilation of the corpus is described in the TSD2011 paper Ljubešić, N., Erjavec, T. *hrWaC and slWac: Compiling Web Corpora for Croatian and*

Slovene. The morphosyntactically annotated and lemmatized corpus is distributed under the CC-BY-SA licence. The first version is freely accessible for querying at http://faust.ffzg.hr/bonito2/run.cgi/first_form?corpname=slwac. A new crawl with an updated crawler is scheduled for 2012-09. The target size of the second version of slWaC is 1 billion words.

3.13. Croatian-English Parallel Web Corpus (hrenWaC)

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

- development of web crawler for parallel hr-en texts
- adaptation of automatic boilerplate removal for parallel texts
- ca 33,000 parallel hr-en texts crawled from top-level internet domain .hr (ca 253,000 translation units)
- automatic alignment on document level
- automatic measuring of parallelity of documents
- manual checking of parallelity of texts
- automatic alignment at sentence level
- manual checking of sentence alignment
- yielded 99,001 aligned translation units
- TMX and TXT/Moses formatting
- the corpus webpage has been produced (http://www.nljubesic.net/resources/corpora/hrenwac/)
- documentation, compilation of metadata and corpus description

The Croatian-English Parallel Web Corpus is a collection of paraellel Croatian-English texts crawled from .hr domain. This corpus was automatically collected by finding on-line documents in English that parallel to the documents already crawled in hrWaC. The parallelity of texts was calculated and selection treshold empirically set to 0.52 on a scale between 0 and 1. After that, the collection of parallel-text candidates has been manually inspected for real parallel texts. The initial crawled corpus had ca 253,000 sentence/translation units pairs (ca 8 Mw per language), while the manual checking resulted in 99,001 sentence/translation units pairs. The corpus is distributed under the CC-BY-SA licence.

3.14. South-East European Parallel Corpus (SETimes Corpus)

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

- the whole corpus was recrawled using stricter extraction process – no HTML residues present
- language identification on every non-English document has been done – non-English online documents contain English material in case the article was not translated into that language

- encoding issues in Croatian and Serbian were resolved – diacritics were partially lost in the first version of the SETimes corpus due to encoding errors – text was rediacritized
- automatic sentence alignment
- the corpus webpage has been produced (http://www.nljubesic.net/resources/corpora/setimes/)
- documentation, compilation of metadata and corpus description

The SouthEast European Parallel Corpus (SETimes Corpus) is based on the content published on the SETimes.com news portal. The news portal publishes "news and views from Southeast Europe" in ten languages: Albanian, Bosnian, Bulgarian, Croatian, English, Greek, Macedonian, Romanian, Serbian and Turkish. This version of the corpus tries to solve the issues present in an older version of the corpus (published inside OPUS, described in the LREC 2010 paper by Francis M. Tyers and Murat Serdar Alperen). The sentence-aligned language pairs are freely downloadable in TMX or TXT/Moses format. The corpus is published under the CC-BY-SA license.

3.15. Croatian Dependency Treebank (HOBS)

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

- manual annotation of ca 3,300 sentences using TrEd editor
- conversion from .fs format to CoNLL format
- adapting of the existing treebank web page (http://hobs.ffzg.hr)
- documentation, compilation of metadata and corpus description

The Croatian Dependency Treebank is a part of the Croatian National Corpus (i.e. Croatian part of the Croatian-English Parallel Corpus, CW2000) where 4,626 sentences (118,529 tokens) are planned to be manually annotated at the analytical layer following the Prague Dependency Treebank formalism adapted to Croatian. The corpus size is currently 3,465 sentences (88,045 tokens). It is published under CC-BY-NC-SA license.

3.16. Web Content Extractor

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

- enhanced algorithm for boilerplate removal and content extraction has been produced
- evaluation of the algorithm has been conducted for Croatian web pages
- the tool webpage has been produced (http://www.nljubesic.net/resources/tools/webcontentextractor/)
- documentation, compilation of metadata and tool description

The Web Content Extractor is a tool for content extraction from web pages for building web corpora. The content extraction algorithm developed for building hrWaC and slWaC is described in TSD2011 paper Ljubešić, N., Erjavec, T. *hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene*. An implementation (a java file) is published under the Apache 2.0 licence. A Croatian evaluation sample used in the paper can also be downloaded and it is distributed under the CC-BY-SA license.

3.17. Collocation and Term Extractor (CollTerm)

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

- calibration of the tool for Croatian
- the tool webpage has been produced (http://www.nljubesic.net/resources/tools/collterm/)
- documentation, compilation of metadata and tool description

The CollTerm is a language independent tool for collocation and term extraction. It is an application that collects collocation and term candidates based on five different co-occurrence measures for multiword units (i.e. collocations) or distributional differences from large representative corpus by application of the TF-IDF measurement on singleword units. The language dependent part consists of stop-word list and list of MWU MSD-patterns that can be coded with regular expressions as well. The application is described in the paper presented at TKE2012 by Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I, Tadić, Gornostay, T. *Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages*. The first version of this application is available as an integral part of ACCURAT Toolkit that is available under Apache 2.0 license (http://www.accurat-project.eu/index.php?p=accurat-toolkit). In this version of the tool a calibration of MWU MSD-patterns has been provided for Croatian thus enhancing the usability of the tool. The plan is to provide calibration for other CESAR languages as well.

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 version 1.1 of the Polish Sejm Corpus contains the following updates:

- new data: transcripts from official parliamentary questions/answers included in the corpus,
- semi-automatic correction of some common typos in the parliamentary session transcripts,
- 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 1.7 of plWordNet contains new portion of data created by semi-automatic extension of previous version and a portion aligned with Princeton WordNet.

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

- Investigation of available standards for representation of aligned wordnets.
- Decision to pursue a Lexical Markup Framework (LMF) based solution.
- Investigation into existing LMF solutions in context of wordnets.
- Adapting to WordNet-LMF schema from EU KYOTO project, with GermaNet modifications.
- Developing a conversion tool from native plWordNet (Słowosieć) XML format to the parallelized LMF format.
- Creating a documentation for native plWordNet format and the parallelized LMF based format.
- Creating the resource description to maintain META-SHARE compliance.

4.6. Polish Named Entity Gazetteer

The Polish Named Entity Gazetteer contains partly inflected entries of Polish (and some foreign) proper names and named entity components (forenames and surnames, geographical names, organizational names, relational adjectives and inhabitant names stemming from country names as well as named entity triggers – months, days, positions, etc.). The resource

was used for the automatic pre-annotation of the National Corpus of Polish (NKJP) at the level of named entities.

This resource in its textual version has been made available in CESAR within the previous batch of language resources. Within the current batch the following work has been carried out:

- An XML format has been defined for the gazetteer in compliance with the LMF standard.
- Conversion and validation tools for the LMF-compliant format have been developed.
- A distribution package, containing a converted and validated XML version, has been prepared. It is divided into two parts:
  - 9,060 fully inflected lexemes and their corresponding 95,359 word forms,
  - 35,884 partly inflected lexemes (mostly foreign names) and their corresponding 40,612 word forms.
- Resource description has been updated.

4.9. TaCo

TaCo is a morphosyntactic tagset converter for positional tagsets.

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

- Standardization of representation of linguistic information: tagset definition based on Spejd formalism, XCES as input-output format.
- Implementation of the converter based on decision trees produced by C5.0 algorithm.
- Cross validation of the TaCo tool on the Corpus of Frequency Dictionary of Contemporary Polish, annotated with National Corpus of Polish tagset and IPIPAN Corpus tagset. Conversion achieved 96.1% of correcntess (= F-measure = weak correctness).
- Tool metadata description created to maintain META-SHARE compliance.

4.10. Spejd

Spejd is a shallow parser, which allows for simultaneous syntactic parsing and morphological disambiguation.

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

- Documentation standardization.
- Tool metadata description created to maintain META-SHARE compliance.

4.11. N-grams from balanced NKJP

The resource contains n-grams of words for n from 1 up to 5 extracted from balanced subcorpus of National Corpus of Polish (http://nkjp.pl). Each word is a maximum series of consecutive non-whitespace characters.

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

- Extraction of plain-text content from the corpus and converting all characters to lower-case.
- Extraction of all required n-grams from the above-mentioned content.
- Sorting the result by the number of unique occurrences.
- Resource description has been created to maintain META-SHARE compliance.

4.12. Redistributable subcorpus of the National Corpus of Polish

The redistributable subcorpus of the National Corpus of Polish consists of all texts of the National Corpus of Polish (see http://nkjp.pl) that are free from intelectual property constraints.

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

- Texts with "free" availability status (marked in publicationStmt/availability element of the TEI header.xml file for each text) have been identified and extracted, preserving the directory structure of the full corpus.
- Corpus metadata descriptions have been created to maintain META-SHARE compliance.

4.13. Morfeusz

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

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

- Morphological data and inflection patterns have been exported from Kuźnia (a system for maintaining morphological dictionaries).
- Resource description has been created to maintain META-SHARE compliance.

4.14. Morfologik Inflectional Dictionary

Morfologik is a comprehensive morphosyntactic dictionary of Polish available under a liberal BSD license.

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

- Improvement of morphosyntactic tagging for all nouns.
- Improving consistency of the tagset.
- Developing a new web tool to maintain the dictionary.
- Joining the resource with SGJP to produce a new resource, PoliMorf, whose Morfologik will be a special version.
- Creating the resource description to maintain META-SHARE compliance.

4.15. The Grammatical Lexicon of Polish Phraseology (SEJF)

The Grammatical Lexicon of Polish Phraseology (SEJF) has been developed within the ERDF Nekst programme and contains 5,000 multiword-word nominal, adjectival and

adverbial lexemes of the Polish general language. Their morphosyntax is described by 160 graph-based inflection paradigms, which allow an automatic generation of about 93,000 inflectional and syntactic variants.

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

- Making SEJF available under the 2-clause BSD licence (FreeBSD) has been successfully negotiated with resource owners.
- The lexicon has been proofread and its extensional form (containing all inflectional and syntactic variants) has been generated.
- A first version of a distribution package has been prepared. It contains textual versions of both the intentional (lexeme-oriented), and the extensional (variant-oriented) lexicon, as well as the corresponding inflection paradigms.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.16. The Grammatical Lexicon of Polish Economical Phraseology (SEJFEK)

The Grammatical Lexicon of Polish Economical Phraseology (SEJF) has been developed within the ERDF Nekst programme and contains over 11,000 multiword-word lexemes of the Polish economical terminology. Their morphosyntax is described by over 300 graph-based inflection paradigms, which allow an automatic generation of about 147,000 inflectional and syntactic variants.

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

- Making SEJFEK available under the 2-clause BSD licence (FreeBSD) has been successfully negotiated with resource owners.
- The lexicon has been proofread and its extensional form (containing all inflectional and syntactic variants) has been generated.
- A first version of a distribution package has been prepared. It contains the textual versions of both the intentional (lexeme-oriented), and the extensional (variant-oriented) lexicon, as well as the corresponding inflection paradigms.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.17. The Grammatical Lexicon of Warsaw Urban Proper Names (SAWA)

The Grammatical Lexicon of Warsaw Urban Proper Names (SAWA) has been developed within the LUNA European project and contains about 9,000 names related to the Warsaw transportation system (streets, squares, buildings, bus stops, etc.). Their morphosyntax is described by over 450 graph-based inflection paradigms, which allow an automatic generation of about 300,000 inflectional and syntactic variants.

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

- Making SAWA available under the 2-clause BSD licence (FreeBSD) has been successfully negotiated with resource owners.

- The extensional form (containing all inflectional and syntactic variants) of the lexicon has been generated.
- A first version of a distribution package has been prepared. It contains the textual versions of both the intentional (lexeme-oriented), and the extensional (variant-oriented) lexicon, as well as the corresponding inflection paradigms.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.18. The multilingual lexicon of toponyms (WikiTopoPl)

The multilingual lexicon of toponyms (WikiTopoPl) contains a list of over 155,000 Polish geographical proper names (countries, cities, regions, hydronyms, etc) and their equivalents in Bulgarian, German, modern Greek, English and Romanian. These data (whenever available) have been automatically extracted from the open encyclopedia Wikipedia. The Wikipedia categories attached to the lexicon entries have been mapped to a short list of succinct categories compliant with Prolexbase, a multilingual ontology of proper names.

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

- Making WikiTopoPl available under the CC-BY-SA 3.0 Unported license (the same as for Wikipedia itself) has been successfully negotiated with resource owners.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.19. Valence dictionary of Polish

The new valence dictionary of Polish verbs constitutes a description of valence of 1433 items, including information about the aspect of the entries, control relationships between arguments, and the status of relevant arguments as subjects or passivisable objects (also among non-nominal complements).

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

- Preparation of a dictionary format suitable for use by both existing parsers (Świgra and the Polish LFG grammar) and future projects.
- Creating a web tool allowing manual edition of the valence frames.
- Automatic conversion of 1433 entries taken from the electronic version of Świdziński's valence dictionary to the established format.
- Using the designed tool to manually correct the errors in automatic conversion and provide missing information.
- Generating a valence dictionary in plain text format.
- Resource description has been created to maintain META-SHARE compliance.

4.20. Summarizer

Summarizer is a tool for creating extractive text summaries.

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

- IPR issues has been clarified and the resources made available under CC-BY licence.
- Resource description has been created to maintain META-SHARE compliance.

4.21. Morfologik-stemming

Morfologik-stemming is a library featuring morphological analysis, spelling correction, and building of finite-state automata for these purposes. It is bundled with a morphological dictionary for Polish, Morfologik.

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

- Morphological data and inflection patterns have been exported from Kuźnia (a system for maintaining morphological dictionaries).
- Resource description has been created to maintain META-SHARE compliance.

4.22. Corpus of the Polish language of the 1960s

The Corpus of the Polish language of the 1960s (originally: the corpus of frequency dictionary of contemporary Polish) was prepared to create a general frequency dictionary of contemporary Polish.

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

- Modifications in Anotatornia (tool for the manual on-line annotation of corpora) to meet the requirements of the corpus.
- Manual annotation of the corpus texts (segmentation and morphosyntactic level).
- Validation and packaging as a set of TEI P5-compatible XML files.
- Resource description has been created to maintain META-SHARE compliance.

4.23. Shallow Grammar of the National Corpus of Polish

Shallow grammar of the National Corpus of Polish is a set of rules which was used for the automatic pre-annotation of the National Corpus of Polish at the syntactic level. It was constructed manually and encoded in the shallow parsing system Spejd (http://nlp.ipipan.waw.pl/Spejd/). It consists of rules for multiword entities, abbreviations, syntactic words, and syntactic groups.

Within CESAR the grametadata description has been created to maintain META-SHARE compliance.

4.24. Pantera

Pantera is a Brill Tagger for morphologically rich languages.

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

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

- Resource description created to maintain META-SHARE compliance.

4.25. PolNet

PolNet is a WordNet like lexical data base built from scratch according to the "merge model" methodology. Its design started in 2006 and continues. The resource development procedure is based on the exploration of good traditional dictionaries of Polish and language corpora investigations (IPI PAN Corpus and domain/application oriented corpora). The PolNet development was organized in an incremental way, starting with general and frequently used vocabulary. We selected the most frequent words found in a reference corpus of Polish language with however one important exception made for methodological reasons. The reason was that we assumed possibly early validation of the resource in a real-size application for which an application-complete vocabulary was necessary.

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.

5. Ulodz resources

5.1. PELCRA parallel corpus collection

General actions for the parallel corpora:

- 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, parser 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.

In addition to resources delivered in Batch 1, the PELCRA parallel corpus collection consists of 3 new resources, detailed below:

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

5.1.1. Multilingual (Polish-) parallel corpora*

The collection consists of news releases of 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. The web crawler used to acquire the EuroParl and ESO data, WebLign, has been made available as a separate resource. Polish texts and their equivalents in all other 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 ESO 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.

5.1.2. OSW Polish-English corpus

The OSW Corpus is a new parallel resource for Polish aqcuired 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.

5.1.3. 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 developed to represent all 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 time-aligned conversational spoken corpus of Polish

This resource is a large 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 an 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 for storing, processing and managing 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.

5.3. 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 became necessary during the alignment and crosslinking of Polish-English resources. The creation of the lemmatizer involved 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, extensive testing was performed using the PELCRA corpora, which helped to eliminate more of the errors present in the BNC data and to judge the performance of the dictionary with varied 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 entitity 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. The first version of the tool, made avaialable by the PELCRA team at the University of Łodź under the GPL license, supports binary classification scenarios which enable the user to detect one of two possible languages, and includes models for distinguishing between Polish and English. 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 (http://www.natcorp.ox.ac.uk/) 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, which were then formatted and imported to the machine learning environment Weka (http://www.cs.waikato.ac.nz/ml/weka/). A Java Application Programming Interface (API) was developed to use the models for language detection. The software was tested through JUnit and integration tests. Final steps included the preparation of a documentation of usage, 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 navigational structure of the respective websites to extract the relevant data aligned on a text level, 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 by the PELCRA team under the GPL license.

6. UBG resources

6.1. Serbian Wordnet

This resource delivered in Batch 1 has been upgraded. These are the amendments performed :

- Serbian Wordnet was enhanced by 650 new synsets. The considerable part of added synsets belong to the Wordnet Affect - synsets corresponding to six emotions: anger, disgust, fear, joy, sadness, surprise). For more details see http://wndomains.fbk.eu/wnaffect.html.
- The relation of Serbian Wordnet to Princeton Wordnet for English was upgraded from PWN 2.0to PWN 3.0.
- For the next release, Serbian Wordnet will be enhanced in size, in such a way to cover all emotions frrm the Wordnet Affect. A web tool will be provided to access it which will enable cooperative work on its development.
- This resource is available for download under license MS NC-NoReD.

6.2. Corpus of Contemporary Serbian

This resource delivered in Batch 1 has been upgraded. These are the amendments performed :

- The contents of SrpKor is updated with new texts in order to increase the size of corpus and to correct the proportions of the text domains. Most new texts are fiction, the rest are general and scientific texts. Newly added fiction includes mostly modern Serbian translations of classic and popular literature (Milne, Poe, Dostoyevsky, Hašek, etc.), but also texts written by Serbian authors in 20th and 21th century (Uskoković, Pekić, Bulatović, etc.). General texts include newspaper and magazine columns ("Politika", "Republika", "Mostovi"), feuilletons ("Večernje novosti") and free electronic books downloaded from the multimedia portal "Peščanik". There are also a few scientific texts, mostly university textbooks and teaching materials (librarianship, philosophy). With a few exceptions, all texts added to SrpKor are valid XML documents compliant toTEI Lite Guidelines.
- Also, the search interface is enriched with several functionalities, mostly filters, enabling the user to reduce the search space by requesting a specific author, text domain, year of publication. The texts to be searched can be original Serbian texts only, or Serbian translations only, or both. There are also morphological filters utilizing the fact that the corpus is tagged with lemma and PoS information.
- The resource is available through web interafce under license CC-BY-NC.

6.4. French-Serbian Aligned Corpus

This resource delivered in Batch 1 has been upgraded. These are the amendments performed:

- The alignment of several texts delivered in Batch 1 was corrected.
- Several new literary and newspaper texts were added to the aligned corpus - approximately 100K words.
- The web interface was improved.

- For the next release this corpus will be enhanced in size and the alignment will be improved
- 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"

This resource delivered in Batch 1 has been upgraded. These are the amendments performed :

- One new language was added in this collection: Slovak; all Cesar languages are now represented.
- 15 new bi-texts were added; there is a bi-text now for all Cesar langauge pairs.
- HTML files of all bi-texts are added now for better visualization.
- For the next release this corpus will be enhanced by addition of new languages - Swedish and Turkish. Bi-texts will produced in the alternative XLiFF format as well.
- Resource is available for download under license MS NC-NoReD.

6.7. English-Serbian Aligned Corpus

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

- Texts and their translations were collected. This corpus consists of English texts translated to Serbian, and vice versa, Serbian texts translated to English. They belong to various domains: fiction, general news, scientific journals, web journalism, health, low, education, movies sub-titles. Corpus contains several texts from the *Acquis* communautaire corpus. The majority of bi-texts were until now used only locally (with exception of texts from the 'Intera' corpus).
- The alignment of all English-Serbian bi-texts were corrected and new bi-texts in TMX format were produced. For the majority of texts the alignment is now one to one on sub-sentence level. For all texts alignment was manually checked.
- This corpus is made available for web search for the first time. For search the same interface is used as for the French-Serbian corpus.
- For the next release this corpus will be enhanced in size, the alignment will be improved, the alternative new XLiFF will be produced.
- The resource is available through web interface under license CC-BY-NC.

6.8. Serbian Nooj module

Serbian NooJ module (SrpNooJ) was produced in the scope of the EU-funded CESAR project. It consists of a set of resources in both alphabets that are in use for Serbian: Cyrillic and Latin. Each set consists of:

- the dictionary properties' definition file (metadata),
- one text – a novel "Dva carstva" (Two empires) from a Serbian author Branimir Ćosić comprising of 86979 word forms,
- a sample dictionary in readable form with 35 lemmas that belong to 8 grammatical classes, with examples of multiword units and derivational morphology,
- a sample of morphological grammars used for lemmas from a sample dictionary – three for simple nouns, two for adjectives, two for verbs, and one for a multiunit noun,

- a syntactic grammar for recognition of one class of named entities – full personal names with their roles or functions,
- a full compiled dictionary (divided in three files: nouns, verbs, and other). It comprises of 85,868 entries: nouns (40,886), adjectives (25,558), verbs (15,366), and other (4,058).

SrpNooJ is available for download from NooJ resorces web page
http://www.nooj4nlp.net/pages/resources.html

6.9. Serbian morphological e-dictionary

Morphological electronic dictionary of Serbian (Ekavian pronunciation) (SrpMD) released in the scope of the EU-funded CESAR project is a version of morpholpogical dictionary of Serbian used in the Nooj corpus processing system and consituting the part of the Serbian Nooj Module (see section 6.8). This version is compiant to MULTEXT-East morphosyntactic specification for Serbian (http://nl.ijs.si/ME/V4/msd/html/msd-sr.html) (with one small deviation form – see section 6.10). It comprises of 3,630,613 entries for 85,721 lemmas covering 11 PoS: nouns (646,867/40,425), adjectives (2,315,640/25,826), verbs (654,159/15,359), adverbs (3233), numerals(4,794/175), conjunctions (83), interjections (218), prepositions (169), pronouns (5,321/104), particles (103), abbreviations (26).

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

6.10. MSD tagged Serbian version of Verne's Novel "Around the World in 80 Days"

The Serbian version of Jules Verne's novel "Around the world in 80 days" has been automatically tagged and manually disambiguated. Serbian morphological e-dictionaries were used for tagging. This was done in the scope of the SEE-ERA project "Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages" (2007-2008). In the scope of the Cesar project the set of morphosyntactic tags used by Serbian e-dictionary were automatically translated to tags conforming to MULTEXT-East morphosyntactic specification for Serbian (http://nl.ijs.si/ME/V4/msd/html/msd-sr.html). There is only a small deviation from this specification concerning the small numbers 2, 3, and 4 (tag 'c' for grammatical number for nouns, adjectives and some forms of verbs).

The final file is compiant to TEI P4 markup of linguistically annotated text. Text structure is also tagged: divisions, paragraphs, and segments (sentences). A short TEI header is provided.

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

For the next batch the new version with MWUs tagged will be provided.

6.11. Bibliša - A tool for enhanced search of multilingual digital libraries of e-journals

This tool is a web application for search of digital libraries of articles from bilingual e-journals in the form of TMX documents, as well as for development of new bilingual lexical resources based on this search. It is founded upon previously developed components for *LeXimir* (work station for lexical resources) and *VebRanka* (web query expansion tool) and uses various lexical resources: Wordnets, e-dictionaries and terminological lists. *Bibliša* can expand search queries both morphologically and semantically, as well as to another language, based on available resources. Presently, it is being implemented for the Serbian/English bilingual e-journal *Infotheca,* which uses Serbian morphological e-dictionaries, Serbian and English wordnets connected via the interlingual index, and a bilingual Dictionary of Librarianship. If the search results reveal a shortcoming in existing bilingual resources, an entry to the new bilingual resource will be initiated.

6.12. Ebart corpus

This corpus consists of texts from the Ebart newspaper archive (http://www.arhiv.rs/).

It is based on their e-archive of texts printed in all major Serbian newspapers and magazines in the period from 2004 to 2012. The whole corpus consists of more than 3.3 million articles (6.3 GBytes of data) and it is grouped by the year of publication. For each year there is approximately 700 Mbytes of data. All articles are grouped by major domains: Society, Economy and Finance, Feuilleton, Culture and Entertainment, Media, Communication with readers, Foreign Politics, Internal Politics, and Sport. Each of these domains is further divided into sub-domains.

The corpus can be downloaded from the location: http://www.arhiv.rs/korpus. This corpus is commercially available.

7. IBL resources

7.1. 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
- Compilation of metadata and corpus description
- Standardisation of text samples and corpus structure
- Linguistic preprocessing and annotation
- Maintenance and deployment of programming tools for corpus compilation and annotation
- Upgrading the web search engine
- Updating of the website of the corpus

The Bulgarian National Corpus is a large representative corpus of Bulgarian. BulNC is constantly enlarged and developed. A uniform framework was developed for structuring BulNC, data storage format and description of the texts. BulNC has been substantially expanded and it now contains 979.6 million tokens (during the last six months it has been increased with app. 50%). Currently, written texts comprise 91.11% of the corpus while spoken texts represent 8.89%. Three basic approaches are implemented for collecting new samples for BulNC: use of readily developed collections of texts; manual collection (by means of Internet browsing and downloading texts); and automatic collection (by means of web crawling). All texts are supplied with extensive metadata description compliant with the well established standards. The metadata comprise of 25 fields. A set of tools was developed for extracting the metadata and compiling the corpus description from the markup formats, in a form of a table with 25 columns. The values in the classificatory information columns are limited to a list of predetermined options which ensures a harmonised approach towards the description of the samples.

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). The structure of BulNC is based on the styles and the domains. Each text is 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.

The information in Bulgarian and English on the webpage of the corpus (http://www.ibl.bas.bg/en/BGNC_en.htm) has been updated as well as the list of publications.

The corpus is a pseudocorpus - the proper texts cannot be distributed, only small excerpts are available through the query interface. The text excerpts are offered under META-SHARE NoRedistribution Non-Commercial license for free.

7.2. Bulgarian-X Language Parallel Corpus

Bulgarian-X Language Parallel Corpus (Bul-X-Cor) includes parallel corpora of 33 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
- Compilation of metadata and corpus 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 a website of the corpus

At present, there are 33 parallel corpora in BulNC which are collectively named Bulgarian-X Language Parallel Corpus and which contain 1.9 billion tokens, comprising the biggest parallel corpus of Bulgarian. Languages are not equally represented: the largest parallel corpus is the Bulgarian-English one (280.8 and 283.1 million words for Bulgarian and English respectively); there are 5 other corpora between 100 and 200 million tokens per language, 16 parallel corpora of size in the range 30-52 million tokens per language, further 7 in the range 1-10 million tokens, and the rest are below 1 million, with the smallest corpora being the Chinese, Japanese and Icelandic with less than 50,000 tokens per language. Each parallel subcorpus within Bul-X-Cor mirrors the structure of BulNC.

All Bulgarian texts in BulNC and English texts in Bul-X-Cor are supplied with extensive metadata description, compliant with the well established standards. The Bulgarian-English parallel corpus is supplied as well with annotation on various levels, while the annotation of other languages has just started.

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 being established – for development and debugging, building test environments, documenting source code, creating repositories of programs.

New information and description of Bul-X-Cor is to be added to the website of the Bulgarian National Corpus.

The corpus is a pseudocorpus - the proper texts cannot be distributed, only small excerpts are available through the query interface. The text excerpts are offered under META-SHARE NoRedistribution Non-Commercial license for free.

7.3. Bulgarian-X Language Parallel Corpus Collocation service

The Collocations service is a web service for search of collocations and different types of statistics over the Bulgarian-X Language Parallel Corpus.
The following work has been carried out for the resource within the CESAR project:

- Development of a collocation service
- Conversion of the corpus format
- Clearance of license terms

The Collocation service employs the free of charge NoSketchEngine, a system for corpora processing that combines Manatee and Bonito. The 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 has been performed. The corpus format has been converted to a format readable by the NoSketchEngine indexing machine.
The results are offered under META-SHARE NoRedistribution Non-Commercial license for free.

7.4. 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 2,000 (November figures) new synsets. To define the scope of the enlargement, the synsets in PWN that lack equivalent synsets in BulNet were automatically extracted, embracing various thematic domains, such as psychology, chemistry, medicine, biology, among others.

- Automatic WordNet expansion using parallel corpora.

To detect automatically the new synset candidates, a knowledge based WSD algorithm is applied to the English part of the Bulgarian-English parallel corpus. After a word alignment the Bulgarian words are associated with the corresponding English synsets without Bulgarian equivalents. Definitions for the newly added synsets are manually compiled by experts. In

most cases, definitions are extensive, covering basic encyclopedic knowledge. 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 – the verbal aspect, the 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.

7.5. 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.

7.6. 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.

7.7. Hydra - tool for developing wordnets

Hydra is a tool for editing, viewing, searching and validating wordnet.
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

The Hydra API for wordnet processing uses abstract language independent of the data representation, the tool supports a multiple-user concurrent access for editing and browsing arbitrary number of monolingual wordnets, it optimizes data visualization as well as enhances editing, undo/redo functions, etc. The search engine works with the wordnet modal language.

The language abstracts the internal data representation and is expressive for the most of the tasks in processing wordnets. Provided that a given wordnet property is definable as a formula in the modal language, the tool determines all the objects in the wordnet structure validating the formula, and hence the property, covering an automatic consistency validation. As a platform-independent system, Hydra has been successfully tested under Linux and Windows. The existing body of code was restructuring so that 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.

7.8. Chooser - annotation tool

Chooser is an OS independent multi-functional system for linguistic annotation, adaptable to different annotation schemata.
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 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. It's compatible with Hydra.
Chooser reuses some of the Hydra's modules. It was corrected to be compatible with the new version of Hydra.
The source code has been made available under the GPLv3 license.

7.9. Bulgarian Sentence splitter and Tokenizer

The Sentence splitter marks the sentence boundaries and the Tokenizer marks string os 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, the regular rules and the 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 a 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 the empty symbols. Regular patterns are used to recognize some simple cases of named entities, that means dates, fractions, emails, internet addresses, abbreviations, etc. The Tokenizer classifies each recognized token (for example: small Cyrillic letters, capital Latin letters, etc.). It utilizes finite state transducers for token recognition and type matching.
The resource is distributed under META-SHARE NoRedistribution Non-Commercial license for free.

7.10. Web based infrastructure for Bulgarian data processing

The web based infrastructure combines Bulgarian tokenizer, sentence splitter, tagger and lemmatizer.
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 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. These features are words, word bigrams and trigrams within a window of words around the currently tagged word; POS tags, POS tag bigrams and trigrams in the current window, and information about suffixes, prefixes, capitalization, hyphenation etc. for the unknown words.

The tagger is trained and tested on manually POS disambiguated corpus. The strategy chosen for training Bulgarian tagger is two passes in both directions; a window of five tokens, the currently tagged word being on the second position; two and three-grams of words or tags or ambiguity classes, lexical parameters as prefixes, suffixes, sentence borders, and capital letters. 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 integration with the lower levels of annotation, full Unicode support and improves the model loading speed.

The Bulgarian lemmatizer determines for a given word form its lemma and detailed morphosyntactic annotation. The lemmatization is based on an unambiguous association between the tagger output and information encoded in a large grammatical dictionary of Bulgarian language. At the tagging, a reduced tagset is used (75 word classes compering to 1029 unique grammatical tags in the dictionary), compiled in a way that the minimum necessary information for unambiguous association with the respective lemma is to be ensured. A small number of rules and preferences are also implemented to limit the ambiguity in lemmatization.

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

9. LSIL resources

9.1. Slovak National Corpus

The following work has been carried out for the resource within CESAR since the inclusion of the corpus in the 1st batch of resources:

- additional texts were included in the corpus
- bibliography and style-genre annotation (metadata) has been separately formatted and published
- corpus metadata have been converted into TEI header format

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 proper texts cannot be distributed.

The metadata are publicly available under the Open Database License v1.0: http://opendatacommons.org/licenses/odbl/1.0/.

9.2. Corpus of Spoken Slovak

The following work has been carried out for the resource within CESAR since the inclusion of the corpus in the 1st batch of resources:

- new recordings have been included in the corpus (increased the size by 60%), up to the current size of 2.6 million tokens
- new, user friendly description of events have been provided in the query results

The database of the Corpus of Spoken Slovak contains audio records of spontaneous and semi-prepared speech from the entire Slovak territory and their text transcripts. Specific characteristics of spoken language are selectively captured in the transcripts, such as irregular structure of an utterance, pronunciation variants, means of speech modulation, and the presence of non-linguistic elements. The Corpus of Spoken Slovak provides material for research and description of the real form of contemporary standard spoken Slovak.

This corpus has been released under following licences (multiple licensing): GNU Free Documentation License version 1.3, Affero General Public License version 3, Creative Commons Attribution – ShareAlike 3.0 Unported License.

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 1st 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 database has been released under the licences: GNU Free Documentation License version 1.3, Affero General Public License version 3, Creative Commons Attribution – ShareAlike 3.0 Unported License.

9.6. Slovak Treebank

This resource has been made available with the contribution of the EuroMatrixPlus project. The Slovak Treebank is a syntactically annotated Slovak language corpus, compatible with the Prague Dependency Treebank. The corpus contains 50 thousand manually syntactically annotated (on the analytical level) sentences, majority of which are annotated by two independent annotators.

9.7. Balanced Slovak Corpus

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

- the current size of 247 million tokens
- the corpus has been separated as a standalone resource (it has been previously available as a subcorpus of the Slovak National Corpus), to be compatible with other resources

This is a balanced corpus, consisting of one third of fictional, one third of informational and one third of professional texts (scientific publications, including popular science, textbooks, manuals, specialized journals). This corpus is considered to be a representative source of data of modern written Slovak.

9.8. Manually Annotated Slovak Corpus

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

- several common error types have been corrected in the corpus
- the corpus has been separated as a standalone resource (it has been previously available as a subcorpus of the Slovak National Corpus), to be compatible with other resources

This is a manually morphologically annotated corpus, using the tagset of the Slovak Morphology Database. The corpus composition is 44.3% fictional, 36.7% journalistic and 19.0% professional texts. The size of the corpus is 1.2 million tokens (punctuation included).

9.9. Language model prim-5.0-sane

The Language model prim-5.0-sane is a language model from the Slovak National Corpus. This model is a 733 million token collection. It is in the iARPA format, using witten-bell smoothing. It was created by the IRSTLM Tooklit. It is lowercased. The model has been released with the contribution of the EuroMatrixPlus project.

9.10. Language model prim-5.0-inf

This language model is of journalistic style. The model is built on corpus of 515 million tokens. The language model is in iARPA format, using witten-bell smoothing. It was created by the IRSTLM Tooklit. The model is lowercased. It has been released with the contribution of the EuroMatrixPlus project.

9.11. Language model prim-5.0-vyv

This language model is based on balanced language. The model is built on the balanced Slovak corpus of 247 million tokens. The language model is in iARPA format, using witten-bell smoothing. It was created by the IRSTLM Tooklit. The model is lowercased. It has been released with the contribution of the EuroMatrixPlus project.

9.12. Corpus of Legal Texts

The corpus has been prepared in collaboration with the Ministry of Justice of the Slovak Republic. It is comprised of legal regulations and other available legal documents (laws, decrees, announcements, directives, protocols, etc.). The size of the corpus is 146 million tokens.

9.13. Slovak Web Corpus

The following work has been carried out for the resource within CESAR since the inclusion of the corpus in the 1st batch of resources:

- additional texts have been included in the corpus (increased the size by 10%), up to current size of 1 thousand million tokens

The Slovak Web Corpus is a corpus of up-to-date Slovak language texts published on the Internet. The corpus is automatically lemmatised and MSD tagged.