ICTPSP
ICT POLICY SUPPORT PROGRAMME

CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

CESAR

Central and South-East European Resources
Project no. 271022

Deliverable D3.1.-B
First batch of language resources: actions on resources

Version No. 1.1
15/12/2011

Document Information

Deliverable number:	D3.1.-B
Deliverable title:	First batch of language resources: actions on resources
Due date of deliverable:	12/12/2011
Actual submission date of deliverable:	15/12/2011
Main Author(s):	György Szaszák (BME-TMIT)
Participants:	Mátyás Bartalis (BME-TMIT) Łukasz Degórski (IPIPAN) Łukasz Dróżdż (ULodz) Radovan Garabík (LSIL) Svetla Koeva (IBL) Cvetana Krstev (UBG) Michał Lenart (IPIPAN) Małgorzata Marciniak (IPIPAN) Maciej Ogrodniczuk (IPIPAN) Gábor Olaszy (BME-TMIT) Piotr Pęzik (ULodz) Adam Przepiórkowski (IPIPAN) Marko Tadić (FFZG) Tamás Váradi (HASRIL) Dániel Varga (HASRIL) Veronika Vincze (HASRIL) Duško Vitas (UBG) Jakub Waszczuk (IPIPAN) Adriána Žáková (LSIL)
Internal reviewer:	RILHAS
Workpackage:	WP3
Workpackage title:	Enhancing language resources
Workpackage leader:	IPIPAN
Dissemination Level:	PP
Version:	1.1
Keywords:	Upload, interoperability, standardization, harmonization, upgrade, extension, linking

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
1.0	12/12/2011	draft	György Szaszák (BME-TMIT)	From all Partners	
1.1	15/12/2011	Final	Tamás Váradi (HASRIL)	finalization	

EXECUTIVE SUMMARY

This deliverable entitled D3.1.-B. is a supplement for deliverable D3.1. (and is also related to deliverable D4.3.), treating first 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 First Batch.

Table of Contents

Abbreviations

Abbreviation	Term/definition
LR	Language Resource
LRT	Language Resources and Tools (either language data or tools)
Partners	Partners of CESAR

Table 1. Abbreviations

0. Scope

0.1. Actions – upgrade resources

Task 3.1 in the DoW is 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).

Although not explicitly mentioned by the DoW under work package 3 (WP3), clearance of license terms and conditions for each individual resource was also a time consuming task in several cases, indispensable for uploading metadata and resources to the META-SHARE network.

For the first upload batch, Partners were primarily focusing on these requirements. However, the preparation of subsequent batches 2 and 3 has also begun and some work related to the extension and linking or cross-language alignment of resources (foreseen rather for batches 2 and 3) was also done. Preparation works are ongoing for almost all resources to be included in next batches.

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 freely process and make available the resources and resource-related materials or the consortium succeeds in reaching an agreement with respective copyright holders.

More details on the selection of resources has already been given in deliverable D2.3.

0.2. Management of the 1st batch

The 1st upload batch and related activities were grouped in CESAR as follows:

- set up META-SHARE node, provide metadata schemes and IPR requirements for CESAR Partners – in close collaboration with META-NET and partner projects –; and pilot service – this action is presented in D4.3.
- metadata description, resource documentation (documentation of the delivery from resource point of view) – presented and listed in D3.1.
- actions on resources as required in WP3 (upgrade, extension, standardization, harmonization, etc.) – presented here in D3.1.-B.

This document is an extension for D3.1. deliverable and its paragraph numbering follows that of D3.1. This means that related activities to a given LRT described under a given section in D3.1. 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 1. If some explanation was found necessary, it is also provided, accompanied by the arguments which justify eventual non-foreseen or otherwise foreseen activities related to the LRT.

Partners of CESAR have not only fulfilled, but even overperformed their obligations, as some resources scheduled for later batches have already been included in batch 1. Partners who have added extra LRTs to batch 1 are planning to either carry on further actions on the already released LRTs – which may contribute to considerably enlarge their usability (be on a higher level of technology, wider range of usability, etc.) – or spend their saved time/costs on exploring new resources or tools to be included into the META-SHARE network.

1. HASRIL resources

1.1. Szeged Corpus 2.0

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

- The website of the corpus has been updated
- Annotation errors have been corrected (concerning lemmas and POS-tags)
- Errors in the XML structure have been corrected
- The documentation of the corpus has been updated
- The documentation has been translated to English
- A new editor has been developed for correcting misspelled words; these are being corrected now for next batch
- The MSD coding system has been revised and MSD codes and lemmas are now being corrected according to the new principles

Updating the website and the documentation will contribute to the easier distribution of the corpus and their English translation will enhance the international usage of the corpus. Error correction was necessary for having more reliable data and the consistency of the annotation can be also improved in this way. The correction of misspelled words will lead to a cleaned version of the corpus texts, which enables to eliminate POS-tagging problems related to unknown words. Moreover, the comparison of the performance of POS-taggers on the original and the cleaned version of the corpus will also become possible. The revision of MSD codes will contribute to the wider usability of the corpus: for instance, more detailed derivational information is encoded (e.g. causation or modality in the case of verbs), which can be exploited in higher level applications such as modality detection or semantic role labeling.

1.2. Szeged Treebank 2.0

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

- The website of the corpus has been updated
- Annotation errors have been corrected (concerning lemmas, POS-tags and syntactic tags)
- Errors in the XML structure have been corrected
- The documentation of the treebank has been updated

- The documentation has been translated to English
- A new editor has been developed for correcting misspelled words and they are being corrected now
- The MSD coding system has been revised and MSD codes and lemmas are now being corrected according to the new principles

Updating the website and the documentation will contribute to the easier distribution of the treebank 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. The correction of misspelled words will lead to a cleaned version of the corpus texts, which enables to eliminate POS-tagging problems related to unknown words. The more accurate POS-tagging in turn may lead to a more accurate parsing. Moreover, the comparison of the performance of POS-taggers and syntactic parsers on the original and the cleaned version of the corpus will also become possible. The revision of MSD codes will contribute to the wider usability of the corpus: for instance, more detailed derivational information is encoded (e.g. causation or modality in the case of verbs), which can be exploited in higher level applications such as modality detection or semantic role labeling.

1.3. Szeged Named Entity Recognition Corpus

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

- The website of the corpus has been updated
- 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.

1.4. Hungarian WordNet

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

- The database has been filtered for BCS concepts
- A bunch of annotation errors have been corrected concerning definitions of synsets, format and non-lexicalized synsets

- Negotiation on licensing issues has been continued
- The website of the database has been updated
- The documentation of the corpus has been updated and expanded
- The detailed documentation has been translated to English

Updating the website and the documentation will contribute to the easier distribution of HuWN and their English translation will enhance 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. In order to settle licensing issues as regards the entire database, negotiations have been continued with ELRA. BCS concepts will be freely available, thus their selection has proved necessary.

1.5. Hungarian Webcorpus

Activities completed:

- Bugfixes and minor enhancements to the web-based interface for the frequency dictionary based on the Hungarian Webcorpus. (http://szotar.mokk.bme.hu/szoszablya/searchq.php)
- Improving the documentation for the downloadable, and writing new documentation for the web-based frequency dictionary.
- Metadata collection.

1.6. Hunglish Parallel Corpus Version 2.0

Our long-held goal was to enlarge and improve the Hunglish Parallel Corpus, the largest collection of Hungarian-English parallel text (approx 2 million sentence pairs). Before CESAR began, a data gathering phase was already finished. (Manually collected 188 document pairs for the modern literature subcorpus. Gathered 10000 document pairs from the CELEX webservice for the legal subcorpus.) Under CESAR, we completely re-built the parallel text processing pipeline that we used to construct the corpus. (See the chapter on hunalign.) This project proved to be more extensive than originally envisioned, so we redirected work hours to it from resources where amount of work was originally somewhat over-estimated (huntoken, hunmorph, hunpos, hunpars). The size of the original Hunglish Corpus Version 1.0 was doubled, to approx. 4 million sentence pairs.

Activities completed:

- Normalized the documents for format and encoding with a combination of manual and automatic processing.

- Fine-tuned and debugged the hunalign-harness pipeline to deal with all the data encountered, including obsolete and rare file formats, incorrectly encoded texts, and mismatching texts.
- Completely redesigned filename conventions and directory organization as compared to the V1.0 release of the corpus. Structured data into this organization by a combination of manual and automatic work.
- Documented the corpus, including providing metadata at the level of individual documents.
- Metadata collection.

1.7. morphdb.hu

The morphdb.hu formal morphological description for Hungarian is intended to be used by the hunmorph morphological analyzer. It was not included in the original work plan, but as hunmorph is part of the work package, this synergy made morphdb.hu a natural target for inclusion.

- We made improvements to the documentation and
- provided metadata for META-SHARE.

1.8. hunmorph tool

The hunmorph morphological analysis framework was added as-is to META-SHARE, as it proved to be compatible with META-NET guidelines in its current implementation. All planned activities were already completed for batch 1:

- Metadata collection.
- Testing.
- Minor improvements to the documentation.

1.9. hunalign tool

We constructed the "hunalign-harness" system that encapsulates the aligner into a pipeline that can work with common text file formats and encodings, and detect typical problems such as incorrectly encoded text, mismatching bidocuments, and bidocuments from some language pair other than Hungarian-English. Hunglish Corpus Version 2.0 was constructed using this pipeline (see there for details). Activities completed:

- Design of a new parallel text processing pipeline around hunalign.
- Integrating our existing tools into the pipeline, including:
- document format conversion,
- sentence segmentation,
- stemming,
- sentence alignment,
- language detection,
- document quality filtering,
- sentence quality filtering.
- Workarounds for converting malformed and obsolete html and doc files to raw text.
- Providing metadata for META-SHARE.

1.10. huntoken tool

The huntoken tokenizer and sentence segmenter was added as-is to META-SHARE, as it proved to be compatible with META-NET guidelines in its current implementation. Activities completed:

- Writing a wrapper around huntoken that integrates it into the hunalign-harness parallel text processing pipeline.
- Testing.
- Minor edits to the documentation.
- Providing metadata for META-SHARE.

2. BME-TMIT resources

2.1. Mindentudás Speech Corpus

The raw material of this corpus was built under two earlier projects, but the goals of these projects did not include the publication of the material. This made necessary a large CESAR effort to convert, organize, and document the already available data. This work is in progress, and will be ready for batch 2. A complex negotiation and legal effort was also required to achieve publication rights for META-NET. For batch 1, the following steps were completed:

- Negotiating the distribution rights with the Depositor.
- Working with the META-NET legal team to adopt the Depositor's Agreement and End User Licence Agreement (META-SHARE Commercial NoRedistribution) for our purposes.
- Extracting the audio tracks from the video material.
- Metadata collection for each of the more than 400 audio files of the corpus. (Still in progress, will be ready for V2.)
- Documentation.

2.2. Word Level Speech Database

Actions carried out within CESAR involved:

- Checking and optimizing the waveforms of the word items (silent part- word – silent part)
- unification of the silent periods at the beginning and ending part of the wave file
- Amplitude correction on word waveforms
- Manual marking of sound boundaries and sound symbols on the waveform
- Manual checking of the marked sound boundaries and the given sound symbols
- Software development for generating unified acoustic images from the waveforms, the sound boundary markers and sound symbols
- Automatic preparation of spectrograms of the words containing also the marked sound boundaries and sound symbols
- Automatic preparation of the intensity in time function of the word containing also the marked sound boundaries and sound symbols

- Automatic preparation of the image of the waveform of the word containing also the marked sound boundaries and sound symbols
- Checking the database elements and the unified structure
- Finalizing the word level 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.

2.3. Hungarian BABEL

Actions carried out within CESAR:

- Metadata description
- Extension and enhancement (still ongoing, for next batches)

The Hungarian BABEL speech database has already been available under ELRA license for a decade. Currently it is included in META-SHARE under ELRA license, its metadata description is provided.

In subsequent batches, BABEL's new enhancements are planned to be shared: phoneme based segmentation for the whole set of phonetically rich paragraphs (currently it is 10%), partial prosodic annotation and partial syntactic analysis of the corpus are being prepared. This work is ongoing at the moment.

License issues are planned to be discussed with ELRA (currently ELRA has the exclusive right to distribute BABEL outside Hungary), if possible, BABEL is planned to be made available under META-SHARE licenses in the next batches.

2.4. Hungarian Broadcast News Database

Hungarian Broadcast News Database was finalized by the end of 2005, and hence, the database was quasi ready. This means that actions carried out within CESAR involved mainly:

- Database cleanup, upgrade for interoperability (transcriptions and audio)

- Documentation harmonization
- Clearance of license terms
- Metadata description

Although the resource was almost ready and characterized by a high interoperability (as it was created in an Europe-wide (COST) effort to assess speaker clustering based on audio), a clean-up was performed and audio transcriptions were validated.

Resource is offered under META-SHARE NoRedistribution Non-Commercial license for free, hence it is a derived work of TV public broadcasts, and original broadcasters agreed on licensing only on these (non-commercial / academic use) conditions with an important restriction: the original video may never be re-broadcasted.

2.5. Sound Gesture Database

Sound Gesture Database is a new and unique resource (lexicon of sound gestures). Actions carried out within CESAR:

- Database finalization and cleanup + enhancement
- Splitting utterances into individual speech files
- Database structure cleanup (creation of a directory structure)
- Documentation preparation
- Clearance of license terms
- Metadata description
- Upload the whole LR to a public server

Sound Gesture database is a multisource audio lexicon. Since the items come from multiple sources, including TV broadcast, telephone conversations the audio encoding was set to universal linear 16 bit 44,1 kHz, although gestures derived from telephone conversations will remain band-limited 300-3400 Hz.

The utterances were enhanced for each gesture type by approx 50%. All gestures are stored in individual wav files. Their naming conventions and the directory which contains them refers to the corresponding gesture.

This resource is made available under META-SHARE Commons BY NC SA license.

2.6. Hungarian Speech Emotion Database

Actions carried out within CESAR:

- 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 emotion transcription labels
- Generating and enhancing documentation
- Clearance of license terms
- Metadata description

Hungarian Speech Emotion Database was a raw set of 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 (they were previously transcribed using two (although not disjunct) sets). Each conversation has one associated transcription file covering all utterances within the directory.

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

Resource itself is a derivative from the original telephone conversations, offered under META-SHARE NoRedistribution Non-Commercial license for free, and also under META-SHARE NoRedistribution Commercial license for a fee.

3. FFZG resources

3.1. Croatian National Corpus (HNK)

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

- collected additional 60 Mw for inclusion in the next version
- started with work on automatic TEI header information generation
- adapting the internal system for enlarging and maintaining HNK according to the system used for SNK
- documentation, compilation of metadata and corpus description

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

3.2. Croatian Morphological Lexicon (HML)

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

- Croatian MulTextEast tagset enhanced with additional morphosyntactic categories for some PoS
- additional 12K lemmas processed
- to each new lemma inflectional pattern manually assigned
- word-forms of all new lemmas generated
- the validity of word-forms checked manually
- investigated procedures for automatic enlargement (e.g. automatic generation of names of inhabitants of states, regions, towns; female *nomina agentis*; possesive adjectives from personal names; etc.)
- compilation of metadata and lexicon description

The Croatian Morphological Lexicon is an inflectional lexicon generated automatically by Croatian Inflectional Generator from ca 122,000 lemmas yielding ca 5,000,000 word forms.

The initial set of lemmas was collected from several existing Croatian mono- and bi-lingual dictionaries, while additional entries were collected via corpus or by means of automatic enlargement of the initial list of lemmas. The automatically generated word forms were corrected for known systemic errors, encoded in utf-8 and stored in MulTextEast Lexica format: lemma[TAB]word-form[TAB]MSD. The MSD-tagset is conformant with the MulTextEast v4.0 reccomendations for Croatian language. However, some additions exist: in surnames gender is left unspecified (-), additional subclassification of adverbials has been introduced etc. At the moment the Croatian Morphological Lexicon is a pseudolexicon, accessible only through the Croatian Lemmatisation Server web query interface or php script call that is tailor-made according to user needs.

3.3. Croatian-English Parallel Corpus

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

- TMX formatting from XML
- documentation in the form of the corpus web-page
- compilation of metadata and corpus description
- development of web crawler for parallel hr-en texts
- development of automatic boilerplate removal
- ca 33,000 parallel hr-en texts crawled from top-level internet domain .hr (ca 250 K sentences)
- automatical alignment on document level
- manual checking of parallelity of texts started for inclusion in the enhanced version of the corpus
- ca 600 texts of Croatian translation of Acquis Communautaire were downloaded from Taiex CELEX database
- conversion into XML and validation with JRC-Acquis DTD started

The Croatian-English Parallel Corpus is a parallel unidirectional (hr to en) corpus of contemporary Croatian standard language collected from articles appearing in Croatia Weekly newspapers, published from 1998 to 2000. The corpus samples were obtained in digital form entirely, converted to XML, aligned using Vanilla Aligner, manually checked and stored in TMX format. Additional texts were collected from two other sources, namely, Croatian translations of Acquis Communautaire and crawled parallel hr-en texts from top-level internet domain .hr. The aligned sentences from these two sources will be used in the enhanced version of the corpus, licensing conditions permitting.

3.4. Croatian Lemmatisation server

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

- enhanced version of HML inserted into server
- documentation, compilation of metadata and tool description
- starting to work on turning web interface into a proper web service

The Croatian Lemmatisation Server (CLS) is a web-based service for lemmatisation, POS- and MSD-tagging of Croatian texts. It accepts input in two modes. Through web form mode it accepts direct query allowing lemmas or word-forms as input, giving all word-forms of lemma or all lemmas that a word-form could belong to, respectively. In both cases, the results are accompanied by MSD-tags as well. In the upload mode the CLS expects a verticalised, utf-8 encoded text in contemporary standard Croatian language and returns a zip file with results of the processing of uploaded file. At the moment the limitation of file size is 50,000 tokens, but there is no limitation on number of files. The processing gives all analysis for each token at unigram level, i.e. line in verticalised corpus, regarding the lemma, POS and MSD. The web interface allows user to select the level of processing needed: just lemmatisation, lemmatisation with POS-tagging or lemmatisation with MSD-tagging. POS and MSD tags follow the MulTextEast v4.0 specifications for Croatian with some additions. Upon registration either as academic or commercial user, a php script call tailored according to user's requests can be provided. Also, the existing Croatian Lemmatisation Server will be turned into a web service that will feature lemmatisation and MSD-tagging of verticalised utf-8 encoded Croatian texts including disambiguation.

3.5. Croatian Vallency Dictionary (CROVALLEX)

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

- generation of browsable version out of basic XML format
- documentation on logical structure of CROVALLEX
- compilation of metadata and lexicon description
- automatic collection of additional verbs started
- automatic generation of additional verbs started
- assigning of verbal valencies to new verbs started

The Croatian Valency Lexicon of Verbs, Version 2.0008 (CROVALLEX 2.0008) is an attempt of formal description of valency frames of Croatian verbs. CROVALLEX 2.0008 was developed as the part of the PhD thesis titled *Approaches to the Development of the*

Machine Lexicon for Croatian Language written by Nives Mikelić Preradović and supervised by prof. dr. sc. Damir Boras at the Department of Information Sciences, Faculty of Humanistics and Social Sciences, Zagreb University. The Functional Generative Description (FGD), being developed by Czech linguists Petr Sgall and his collaborators since the 1960s, is used as the background theory in CROVALLEX 2.0008. for the description of valency frames of selected verbs. CROVALLEX 2.0008 contains 1740 verbs that were selected from the Croatian frequency dictionary, according to their number of occurrences. The browsable HTML version of the basic XML format has been produced. Additionally, system for automatic generation of additional verbs, i.e., verbal aspectual pairs, has been produced and it will be used for enhanced version of CROVALLEX.

4. IPIPAN resources

4.1. Polish Sejm Corpus

The Polish Sejm Corpus is a new resource containing collection of stenographic transcripts of Polish Sejm sittings from 1-6 terms of office (1991-2011). They have been automatically tokenized, lemmatized, morphosyntactically described, disambiguated, annotated with syntactic words, groups and named entities and are represented in TEI P5 format.

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

- Session data has been received from Sejm after formal negotiations. Transcripts for terms 5 and 6 were received in the form of DOC and XML files (contrary to publicly available HTML and PDF files) the latter of which contained detailed session structure (publicly not available). Sample audio and video resources have been also acquired.
- Representation format has been developed for non-textual data and TEI P5 format based on NKJP (National Corpus of Polish) representation has been adopted for the textual data.
- Conversion scripts have been developed and run.
- Corpus metadata description in the form of extended TEI header has been prepared.
- AudioText corpus metadata description has been created to maintain META-SHARE compliance.

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

- The morphological data of Morfeusz SGJP (previously available only on AGPL license), Morfologik and several underlying older tools (ispell dictionary and sjp.pl dictionary – the original sources used by Morfologik) have been negotiated with their owners and made available under 2-clause BSD (FreeBSD) licence.
- Cooperation between the maintainers of the dictionaries has been initiated, leading to the creation of a single large morphological dictionary for Polish, comprising and extending both Morfologik and Morfeusz.

- A dedicated tool for extending the dictionary with new lexemes has been specified and developed (code name: Kuźnia, currently in the final stages of development); the tool will allow linguists to add lexemes and their morphological specification in a distributed fashion, over the Internet.
- Tagset converters have been prepared to maintain compatibility between component resources.
- All available morphological data have been imported into Kuźnia and converted into internal representation format.
- Various quality control mechanisms have been implemented, to minimize errors in the resulting dictionary.
- The first version of the new morphological dictionary resulting from the semi-automatic merger of Morfeusz and Morfologik has been exported and made available in the simple 3-value orth-base-tag format.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.3. Polish WordNet

The plWordNet (Słowosieć) is a semantic network which reflects the Polish lexical system. It is under development since 2005 by Wrocław University of Technology and its current version 1.5 has been released recently.

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

- Making the plWordNet available under a liberal license has been successfully negotiated with resource owners.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.4. Polish Named Entity Recognition Tool

Nerf is a tool for Nested Named Entity Recognition based on the Conditional Random Fields modelling technique. It is currently made available.

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

- Making Nerf available under a liberal license has been successfully negotiated with resource owners.

- A distribution package has been prepared.
- New models have been developed and trained.
- Input/output formats have been adjusted.
- The tool has been linked to a rule-based component for simple types (money and percentage expressions) to be able to create hybrid solutions.
- Tool metadata description has been created to maintain META-SHARE compliance.

4.5. 1 million subcorpus of National Corpus of Polish

The 1-million subcorpus of the National Corpus of Polish is manually annotated TEI P5-compatible resource containing multi-level linguistic annotation (segmentation, morphosyntax, syntactic words and groups, named entities, word senses).

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

4.6. Polish Named Entity Resources

The Polish Named Entity Resources (plNER) are currently limited to the Gazetteer for Polish Named Entities, a textual source used within the SProUT platform, which contains 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) on the level of named entities.

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

- Making the plNER available under the 2-clause BSD licence (FreeBSD) has been successfully negotiated with resource owners.
- Copyright-protected inhabitant names and relational adjectives stemming from Polish settlements (owned by the PWN Scientific Publishers) have been removed from the original source.
- A distribution package has been prepared.
- Lexicon metadata description has been created to maintain META-SHARE compliance.

4.7-8. LUNA.PL and LUNA-WOZ.PL corpora

The corpora are transliterated complex spontaneous human-human (LUNA.PL) and human-computer (LUNA-WOZ.PL) dialogues acquired in the course of *Spoken Language UNderstanding in multilinguAl communication systems* (LUNA) project between 2006 and 2009. The source data have been collected at the call centre of the Public Transport Authority of Warsaw and annotated in terms of semantic constituents and semantic structures. They are registered as two separate corpora because of their different nature and description model.

The following work has been carried out for these resources within CESAR:

- The permission to make the data publicly available has been negotiated with the Public Transport Authority of Warsaw and the EXIT Publishing House (which published the previous version of resources).
- Both corpora have been made available under 2-clause BSD (FreeBSD) license.
- TEI P5-based annotation format has been developed by adopting and extending the NKJP format (see http://nlp.ipipan.waw.pl/TEI4NKJP/).
- Original file structure has been mapped into NKJP-compatible file structure with the following annotation layers:
    - audio (wave files),
    - transcription (converted to NKJP text structure layer),
    - words (containing morphosyntactic annotation – split into NKJP segmentation and morphosyntax layers),
    - dialogue turns (integrated into text structure),
    - chunks (syntactic words layer).
- New TEI P5-compatible format has been developed for for the layers currently absent in NKJP:
    - domain attribute-value pairs (concept layer),
    - predicate structure (frames layer),
    - anaphora.
- Conversion of the corpora into the new format has been carried out.
- AudioText corpus metadata descriptions have been created to maintain META-SHARE compliance.

5. Ulodz resources

5.1. PELCRA Polish-English parallel corpora

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.

Specific actions for the PAN Academia Corpus:

This Academia Corpus is a completely new and manually aligned parallel resource for Polish developed entirely within the CESAR project. It contains over 350 000 words of scientific articles published in Polish and English from *Academia – the Magazine of the Polish Academy of Sciences.* The articles were first web-crawled at ULodz and converted from the PDF format and aligned semi-automatically at the sentence level using the memoQ CAT environment. The initial sentence-level alignment was then manually verified and corrected and the texts were further annotated with bibliographic information. The corpus has been made freely available in the TEI P5 and XLiFF formats under the Creative Commons Attribution Non-Commercial license (CC-BY-NC). The Academia Corpus has a formal IPR clearance; the license was negotiated with its publisher and a written permission was obtained on 25th of November 2011 to make it available under CC-BY-NC.

5.2. PELCRA Polish-English parallel corpora

Preparatory work – see 5.1 above.

Available at http://nlp.ipipan.waw.pl/metashare/browse/30/

a) CORDIS and RAPID collections

The CORDIS & RAPID collections of aligned Polish-English texts were web-crawled from the respective websites. The ULodz team developed a dedicated web-crawler in order to acquire this data. The collection contains over 7.5 million words (per one language) of news releases published at http://cordis.europa.eu/ – the Community Research and Development Information Centre in Polish and 5 other EU 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.

b) JRC Acquis

The Polish-English subset of the JRC Acquis collecton was imported into the central RDB database. A dedicated web-crawler was developed to collect additional metadata from the Eur-lex database and integrate it with the collection. Bibliographic information was added to the corpus. TEI P5 and XLiFF version were prepared and made available. Documentation and META-SHARE XML headers were generated. Released in the META-SHARE repository under the Creative Commons Attribution license.

5.3. PELCRA Polish spoken corpus

Available at: http://nlp.ipipan.waw.pl/metashare/browse/32/

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 347 transcriptions, 1.4 million words, over 100 hours of transcriptions of spontaneous conversations in Polish recorded in the years 2000-2011 annotated structurally and bibliographically. 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:

- Enhancement of a TEI P5-compliant schema for spoken transcripts.
- Development of a central RDB system used to store, process and manage the

transcriptions.

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

6. UBG resources

6.1. Serbian Wordnet

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

- Serbian Wordnet was enhanced, from 15,200 synsets to 16,891 synsets. The major part of added synsets belongs to some specific domains, like biology, zoology, chemistry, geology and named entities.
- The existing wordnet was improved and corrected in various ways: many hanging links were removed (either deleted or connected) and new relations among synsets added (mostly of the type holo_part, holo_member and holo_portion).
- The wordnet is made available for the first time for download in XML export format (for non/commercial purposes).
- For the next release, Serbian Wordnet will be enhanced in size, it will be made conformant to the latest Princeton Wordnet, and the set of semantic relations between synstes will be enriched.
- Resource is available under license CC-BY-NC.

6.2. Corpus of Contemporary Serbian

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

- The part of corpus available on web was enhanced – from 23 million words to more than 113 million words.
- The whole corpus was lemmatized and PoS tagged – the previous version did not contain that information.
- All texts in corpus were classified using UDC classification – the previous version did not contain that information.
- The corpus can be searched using the existing web interface which does not allow search using the additional information (items 2 and 3); for the next release the corpus will be enhanced in size and the new more powerful interface will become operational.
- Resource is available under license CC-BY-NC.

6.3. Serbian Lemmatized and PoS tagged corpus

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

- This corpus represents a small portion of the lemmatized and PoS tagged Corpus of Contemporary Serbian (see 6.2). It consists of text for which copyrights were obtained that permit download for non-commercial use. This information is made available for download for the first time.
- For the next release this corpus will be enhanced in size.
- Resource is available under license CC-BY-NC.

6.4. French-Serbian Aligned Corpus

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

- The alignment of all French-Serbian bi-texts – until now available only for local search and use – 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 – until now only one French-Serbian bi-text was uploaded and searchable through web interface.
- For the next release this corpus will be enhanced in size, the alignment will be improved, the alternative new XLiFF will be produced and the new web interface allowing more effective search of bi-texts will become available.
- Resource is available under license CC-BY-NC.

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

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

- The part of this corpus is made available for the first time for download (for non-commercial use). This part of corpus consists of bi-texts involving 17 languages. All texts are aligned to either French original, or Serbian or English translation.
- Two new languages were added for this release: Hungarian and Albanian.
- The alignment of all bi-texts was corrected and for the majority of texts the alignment is now one to one on sub-sentence level. For all texts alignment was manually checked.
- For the next release this corpus will be enhanced by addition of new languages, so

that all Cesar languages will be covered, and some Meta-net languages will be added as well. Moreover, bi-texts for all combinations of Cesar languages will be produced as well as the alternative XLiFF format.

- Resource is available under license CC-BY-NC.

6.6. Organizing Digitized Material

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

- The first version of this software tool was produced for organizing digitized cultural heritage material belonging to ethnographic maps of Serbia. This new version was made independent of actual application.
- The new version was made available for web use and for a new (test) application.
- Resource is available under GPL license.

7. IBL resources

7.1. Bulgarian National Corpus

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

- Increasing the size of the corpus in a balanced way
- 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
- Clearance of license terms

Bulgarian National Corpus (BulNC) is a large representative corpus of Bulgarian. A web-based search application provides access to the corpus.

BulNC is constantly enlarged and developed. A uniform framework was developed for structuring BulNC, data storage format and description of the texts. As of November 2011 BulNC includes 144,669 text samples which comprise of 470,1 million tokens and has been increased with app. 12.5%. 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). Most of the texts downloaded from the Internet are in xml or html format - the raw text is automatically extracted from the markup.

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. The corpus documentation is compiled as a table with 25 columns. A row (a record) corresponds to a single text sample. 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.

Several sets of tools are used for performing various tasks – collection of texts, compiling metadata, linguistic annotation, etc. To ensure easy collaboration, knowledge exchange, code reusability a uniform framework for all programming tools is established – for development and debugging, building test environments, documenting source code, creating repositories of programs.

The Corpora search system has been redesigned. The system uses two different corpora query languages - DQL, designed for Corpora search engine, and MySQL Regular Expressions. A searching server available through http has been developed. The code is refactored such as the searching server to be transparent for the web application. Three different user profiles are designed and along with the not logged user (Guest) they define four types of level access.

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 National Corpus Collocation service

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

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

Collocations service is a web service for collocations search and different types of statistics. The Collocation service employs the free of charge NoSketchEngine, a system for corpora processing that combines Manatee and Bonito. The Collocation service is implemented as a RESTful web service, supporting complicated queries through http. The query returns the collocations of a given word in NoSketchEngine format. The system also supports additional arguments, namely all that are accepted by NoSketchEngine, provided with default values. The service is protected with the HTTP Digest authentication.

New indexing of the Bulgarian National Corpus is performed. The corpus format is converted to a format readable by the NoSketchEngine indexing machine.

The users may observe the frequency of words and language constructions, and generate frequency lists and language models. The results are offered under META-SHARE NoRedistribution Non-Commercial license for free.

7.3. Bulgarian Part-of–Speech Corpus

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

- Conversion of the corpus format
- Development of corpus database
- Implementation of web query engine
- Clearance of license terms

The corpus contains 174,697 lexical units, annotated with the most appropriate grammatical information following the Bulgarian Grammatical Dictionary. The query interface provides searching by word, lemma and combination of the two. The result is a list of available occurrences, their number, gloss and examples extracted from the corpus. The corpus data is converted in a format appropriate for the web application. Flat XML is applied to encode words and annotation (lemma, pos, etc.). Special attention is paid to the text reconstruction, as the punctuation marks were stored as separate words.

UTF-8 encoded database is set to accommodate the corpus data. For every <word, lemma, pos> triple, the database stores number of occurrences and usage examples. A script filters the corpus data, defines the database scheme and fills it with the respective data.

The system is implemented entirely in Python. The web framework is Django. Modules of the multi-purpose annotation tool were redesigned and they are exploited for part-of-speech format management, transformation and searching for particular words and lemmas. The user interface enables efficient online search of language patterns and forms as well.

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

7.4. Bulgarian Sense-annotated Corpus

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

- Sense annotation of lexical units
- Review and correction of already applied annotation
- Annotation upgrade
- Enhancement and standardisation of the annotation
- Documentation extension
- Implementation of web query engine
- Clearance of license terms

Each lexical item (simple word, compound word or multiword expression) is assigned manually the unique semantic or grammatical meaning from the Bulgarian wordnet (BulNet) in the particular context. The Bulgarian Sence-annotated Corpus (BulSemCor) contains app. 99,000 sense-disambiguated lexical items defined in the context of occurrence. The semantic annotation follows the model of the Princeton SemCor, but unlike SemCor each annotated item has been assigned a sense, i.e., both lexical and function words were subject to annotation.

BulSemCor lexical units has been annotated by experts independently. The annotation has been validated and unified to prepare the resource for its end release.

As Bulgarian wordnet (BulNet) has been regularly enriched (new synsets and literals have been added, groups of words have been interpreted as multiword expressions, among others) the annotation of the BulSemCor has to be changed to follow the updated BulNet database. Changes in the database resulting in a lack of correspondence between BulNet and BulSemCor senses were registered automatically. Identified instances of mismatch had to be manually corrected.

Misunderstandings and conflicting interpretation among the annotators were clarified, leading to a final review and correction of the annotated data in the framework of a standardised strategy.

Documentation on the BulSemCor annotation (in Bulgarian) was extended to include specification of the annotation schema, information about the annotation methodology, guidelines and principles of inter-annotator agreement.

An application for BulSemCor online demonstration was built. Users may submit queries to a search engine according to word form, lemma or combination of the two. The output gives number of occurrences for all senses found in the BulSemCor, BulNet explanatory definition and examples for each sense extracted from the semantically annotated corpus.

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

7.5. Bulgarian X language Parallel Corpus

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

- 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
- Clearance of license terms

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 languages are not equally represented. The largest is the English corpus with 80 mln. tokens, there are 21 corpora in the range 30-52 mln. words, 7 in the range 1-10 mln. and the rest contain less than 1 mln. words (data are valid as of November 2011). The automatic collection of corpora was preferred for collecting large amount of parallel texts and for that purpose a crawler was designed. It is then adjusted and optimised for each available source. Web structure mining is employed to reduce the number of visited links and improve efficiency. The raw text is automatically extracted from the xml or html markup.

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 metadata comprise of 25 fields. A set of tools was developed for extracting the metadata and compiling the corpus description from the markup formats. The corpus documentation is compiled as a table with 25 columns. A row (a record) corresponds to a single text sample. 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.

The Bulgarian-English parallel corpus is supplied with annotation on various levels while the annotation of other languages has just started. Apache OpenNLP with a pre-trained model is exploited for the English texts annotation – sentence segmentation, tokenization, POS tagging and syntactic parsing. The lemmatisation of English texts is performed using the RASP lemmatizer. Some work was carried on to ensure interoperability between linguistic annotation tools and uniformity of linguistic annotation between languages – the work includes some changes of annotation formats, unification of categories and mapping of tagsets for English and Bulgarian texts. The texts comprising the Bulgarian-English parallel corpus are also aligned at sentence level.

Several sets of tools are used for performing various tasks – collection of texts, compiling metadata, linguistic annotation, etc. To ensure easy collaboration, knowledge exchange, code reusability a uniform framework for all programming tools is established – for development and debugging, building test environments, documenting source code, creating repositories of programs.

The Corpora search system is enhanced to support parallel corpora in a uniform way. For a given query the system retrieves matches in all documents irrespectively of the language. To achieve this several actions were performed: establishment of a new file naming strategy, enabling web interface with aligned sentences visualisation; implementation of aligned sentence retrieval, modification of indexing machine to serve, modification of the database feeding script.

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.

7.6. Bulgarian wordnet

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

- Alignment with Princeton wordnet 3.0
- Enlargement of Bulgarian wordnet with new synsets, literals and relations
- Review and correction of existing synsets, literals and relations
- Clearance of license terms

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 Bulgarian wordnet was aligned with the Princeton wordnet (PWN) 2.0. The BulNet was upgraded automatically to PWN 3.0. The automatic transformation of the source model necessitated subsequent review of the existing synonym sets, their literals, and relations between the synsets. The identification of the synsets that had to be merged and split was carried out automatically followed by a manual editing.

The Bulgarian wordnet was manually enlarged with around 6,000 (November figures) new synsets. The enlargement of the Bulgarian wordnet is closely related to the work on the word-sense annotated corpus of Bulgarian - BulSemCor - where each lexical item is assigned a single wordnet sense. 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. 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 with new relations, namely, hyponyms, meronyms, antonyms, etc. defined between the newly added and already existing synsets. Extension in the number of literals is not only a consequence from the increasing of the wordnet itself but it is due to some specific peculiarities of Bulgarian - verbal aspect, rich derivational system, etc.

Validation 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. At the end of the period, an automatic validation for consistency was carried out followed by manual correction of the identified errors (lacking relations and others).

The latest version of the Bulgarian wordnet is spread by ELDA. The resource is offered under META-SHARE NoRedistribution Non-Commercial license for fee, and under META-SHARE NoRedistribution Commercial license for a fee.

7.7. Wordnet web service

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

- Development of the wordnet database
- Development of the web service
- Clearance of license terms

Wordnet service is an online service that provides an access to a subset of the Bulgarian wordnet (BulNet), containing over 12,00 synonym sets (synsets) from the so called Base Concepts subset 1 and to the entire database of the Princeton Wordnet (PWN). A new database is fed to serve the Wordnet service. As a result users can search for synonyms, hypernyms, antonyms, and translation equivalents of different words and lemmas in the following language pairs: English-English, English-Bulgarian, Bulgarian-English, and Bulgarian-Bulgarian.

The system is RESTful web service and supports two types of queries through http: for objects where the query represented in the Wordnet modal language returns a list of object identifiers; for information related with the objects that returns a list with data for Literal (identifier, word, lemma), Synset (identifier, ili, pos, definition, stamp, bcs, language (identifier), frequency), Note: (identifier, text). The implementation is entirely in Python. Twisted library is used for the web server, with a HTTP Digest authentication protection. Some modules of Hydra (the system for development and validation of wordnet databases) are redesigned for data management, database construction and wordnet queries.

The resource is offered under META-SHARE NoRedistribution Non-Commercial license for free.

7.8. Bulgarian Spell Checker for Windows

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

- Implementation of the Spell Checker engine
- Development of spell checker dictionary
- Identification of candidates for correction
- Implementation of error analysis and suggestion generation module
- Integration with the Microsoft windows API
- Clearance of license terms

The system for spelling checking for Microsoft Office detects and marks the incorrectly written words in a text and suggests the most probable candidates to correct the errors. The spellchecker is realised as 32-bit dynamically linked library (DLL). The spell checker engine performs lookups in a dictionary, identifies misspelled words and generates candidates for a correction. The implementation efficiently exploits finite states automata. The functionality of the product is realised through the use of minimal acyclic deterministic automata and Levenshtein automata, which allow maximum speed, precision and coverage.

The Spell Checker dictionary is based on the Electronic Grammar Dictionary of Bulgarian and it has been extended with more than 10,000 lemmas.

A logic for detection of performance errors (wrong key pressed, letter swapping, skipped letters or extra letters) and competence errors (identified by means of rules) is applied while identifying candidates for a correction.

The Spell Checker can be run under Microsoft Windows XP or Microsoft Windows 7, supports Microsoft Office 2007 / 2010 - 32-bit version and works with the standard BDS as well as the phonetic layouts.

The resource is offered under META-SHARE NoRedistribution Non-Commercial license for free.

7.9. Bulgarian Spell Checker Web Service

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

- Development of the web service
- Integration in various types of web application
- Clearance of license terms

The spell checker engine is integrated as a web service – both the web service integration and the online spelling checking (as an illustration of the integration) are possible. The spell checker engine performs lookups in a dictionary, identifies misspelled words and generates candidates for a correction. The engine is based on the construction of a dictionary in a minimal acyclic deterministic automaton and offers replacement suggestions on the basis of Levenshtein automata.

The web service allows an integration of the spell checker engine in various types of web applications through a Java script component, which provides requests to the service as well as the communication with the user interface. The service is compatible with jquery spellchecker - a Java script component for web page user interface.

The resource is offered under META-SHARE NoRedistribution Non-Commercial license for free.

8. LSIL resources

8.1. Slovak National Corpus

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

- Slovak National Corpus has been cleaned of incorrectly converted texts
- additional texts were included in the corpus (increased the size by 36%), up to final size of 770 million tokens
- new version 5.0 has been released
- filter used to discard foreign language text fragments has been tuned for better accuracy
- existing duplicates have been eliminated
- web interface to the corpus has been provided (using bonito2/NoSketch engine)

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

8.2. Corpus of Spoken Slovak

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

- new recordings have been included in the corpus (increased the size by 140%), up to final size 1.6 Mtokens
- new version 3.0 has been released
- transcription rules have been modified to include additional phenomena and to exclude seldom used tags
- existing transcriptions have been checked for inconsistencies

The database 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.

8.3. Slovak Morphology Database

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

- the database markup has been extended to indicate substandard variants of the word forms
- Slovak Morphological Database is a database of lemmas and their inflected wordforms with MSD tags

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

8.4. Slovak-Czech Parallel Corpus

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

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

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

8.5. Slovak-English Parallel Corpus

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

- the corpus has been extended by Slovak → English translated texts (fiction), in addition to the original English → Slovak direction
- missing bibliography annotation has been added to the corpus entries
- texts have been cleaned, incorrect alignments have been corrected
- a dictionary based interface to query the aligned phrase table constructed out of the corpus has been created and published at the Ľ. Štúr Institute's dictionary portal: http://slovniky.korpus.sk/?d=psken

The corpus consists of parallel Slovak and English texts (mostly fiction), with automatic lemmatization, morphological analysis (for Slovak), POS tagging (for English). The corpus consists of original English language books and their Slovak translations. This is a pseudocorpus, only the query interface is available, the texts proper cannot be distributed.