UDK: 004.9:811.16’374
811.16'322

Central and South-European language resources in META-SHARE

Maciej Ogrodniczuk (maciej.ogrodniczuk@ipipan.waw.pl),
Institute of Computer Science, Polish Academy of Sciences
Radovan Garabík (garabik@kassiopeia.juls.savba.sk),
Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences
Svetla Koeva (svetla@dcl.bas.bg),
Institute for Bulgarian Language, Bulgarian Academy of Sciences
Cvetana Krstev (cvetana@matf.bg.ac.rs),
University of Belgrade, Faculty of Philology
Piotr Pęzik (piotr.pezik@gmail.com),
University of Łódź
Tibor Pintér (pinter.tibor@nytud.mta.hu),
Research Institute for Linguistics, Hungarian Academy of Sciences
Adam Przepiórkowski (adam.przepiorkowski@ipipan.waw.pl),
Institute of Computer Science, Polish Academy of Sciences
György Szaszák (szaszak@tmit.bme.hu),
Dept. of Telecommunications and Media Informatics,
Budapest University of Technology and Economics
Marko Tadić (marko.tadic@ffzg.hr),
University of Zagreb, Faculty of Humanities and Social Sciences
Tamás Váradi (varadi@nytud.hu),
Research Institute for Linguistics, Hungarian Academy of Sciences
Duško Vitas (vitas@matf.bg.ac.rs),
University of Belgrade, Faculty of Mathematics

Abstract

The paper intends to give a brief summary of one the most recent efforts on building the pan-European language technology infrastructure: META-NET – a network of Excellence consisting of 54 research centres from 33 countries – and specifically, its Central and South-European participating project: CESAR. One of the major activities of the project is selection of the resources and tools to be collected, validated, standardized, upgraded/extended/cross-lingually aligned and stored in the META-SHARE open resource exchange facility.

The contribution focuses on presenting the repository maintaining the metadata of the selected resources, the methodology and criteria for their selection and a detailed view to the resources and tools delivered by the project in 2011. After highlighting the concepts of META-SHARE metadata model and synchronized network of metadata servers, the article presents the methodology and criteria for the resource selection by calculating point values basing on solid evaluation indicators such as resource availability, quality, and quantity of similar resources available, coverage, maturity, sustainability and adaptability.

The META-NET Language White Papers – the series of reports on the state of each European language with respect to language technology is also presented as well as the licensing guidelines put forward by the META-SHARE community, promoting open and free of charge use of data and tools by using standardized and well-defined legal attributions.

Keywords
language resources and tools, metadata, resource repository, open linguistic infrastructure, language whitepapers, Slavic languages, Bulgarian, Croatian, Hungarian, Polish, Serbian, Slovak

Introduction

As the goals of the information society are increasingly becoming a reality in our lives, the challenges that language, the inherently human medium of communication, transmitted through whatever technology, are emerging as one of the fundamental issues. Hence the importance of language technology as a key enabling technology that is required in a wide range of domains from breaking down language barriers and preserving cultural heritage to extracting knowledge through text analytics – to cite only a few areas that may be of interest to readers of this journal.

Language technologies crucially depend on language resources (LT), vast amounts of carefully analysed and annotated data, as well as appropriate tools and standardised methods to process these resources. Concern with the widespread availability of data and tools in a standard and well-documented way including clear intellectual property status (IPR) status, has given rise to the concept of language technology infrastructure. One of the most recent pan-European efforts is META-NET, a network of Excellence based on four participating projects, one of which is the CESAR project, which was reported upon in an earlier issue of this journal (Varadi, 2011).

The present article, written by partners in the CESAR project, is intended to give an overview of the META-SHARE infrastructure as well as an account of the initial stock of language resources and tools prepared by the CESAR consortium.

1. META-NET and META-SHARE

META-NET is a Network of Excellence dedicated to fostering the technological foundations of a multilingual European information society. Language Technologies will:

- enable communication and cooperation across languages,
- secure users of any language equal access to information and knowledge,
- build upon and advance functionalities of networked information technology.

A concerted, substantial, continent-wide effort in language technology research and engineering is needed for developing applications that enable automatic translation, multilingual information and knowledge management and content production across all European languages. This effort will also enhance the creation of intuitive language-based interfaces to technology ranging from household electronics, machinery and vehicles to computers and robots.

To this end META-NET is building the Multilingual Europe Technology Alliance (META). Bringing together researchers, commercial technology providers, private and corporate language technology users, language professionals and other information society stakeholders, META will prepare the necessary ambitious joint effort towards furthering language technologies as a means towards realising the vision of a Europe united as one single digital market and information space.

1.1. The metadata model

One of the most valuable META-SHARE features is a global view on resources, going far beyond the repository idea. Apart from providing a catalogue of LRTs (data, tools, technologies) it also intends to maintain information that can be used to enhance their exploitation, such as reference documents (papers describing the resource, associated reports, tagset manuals, guidelines for LR production, etc.), information on persons and organizations involved in their creation and use (e.g. creators of resources, funders, distributors, etc.), links to related projects and activities (e.g. projects that have funded the creation of an LR, where an LR has been exploited, etc.) or licenses (for the distribution of the LRs).

META-SHARE metadata model is based on three concepts from the Component Metadata Infrastructure (Broeder et al. 2008):

- components – containers for elements gathering semantically coherent properties into coarse-grained sets,
- elements – encoding specific descriptive features of the LRs,
- relations – linking together interrelated resources from the META-SHARE repository (e.g. original and derived, raw and annotated resources, a language resource and the tool that has been used to create it, etc.)

According to this principle, a common META-SHARE metadata model was prepared and encoded in XML Schema notation. It offers two basic levels of resource description: the minimal schema with basic description of the resource, retained for compatibility, and the maximal schema with higher degree of granularity, providing more detailed information on each resource. Similarly, element classes were divided into groups according to their importance and dependence: mandatory, condition-dependent mandatory (has to be filled in when specific conditions are met), recommended (LR producers are advised to include information) and optional.

There are two levels in the taxonomy of language resources in the model: first level referring to the resource type (corpus, lexical/conceptual resource – terminological resources, word lists, semantic lexica, ontologies, language description and technology/tool). The second level provides type-dependent subclassification (language, medium, domain, format, annotation features, etc.)

1.2. META-SHARE repository

META-SHARE is the open distributed facility for the sharing and exchange of resources provided by META-NET. META-SHARE servers form a chain of interrelated nodes, offering seamless access to resources for their users and sound editing and administrative interfaces for resource owners.

META-SHARE repositories contain resource descriptions in the form of metadata conformant to the metadata model described above and are populated and updated by means of harvesting data from existing collaborating initiatives and projects, as well as uploading new resource descriptions via metadata editor and maintaining access permissions. As soon as the metadata of a new resource are made available to one of the META-SHARE nodes, the network automatically propagates and synchronizes descriptions among the other registered nodes (see Figure 1).

Figure 1. Result of searching a META-SHARE repository [1] by filtering by language (Polish) and multilinguality type (parallel). Four such resources were find, all of them provided by CESAR.

2. Populating META-SHARE

META-NET and META-SHARE concepts were first tested on a pan-European scale in early December 2011, when all 3 sister ICT-PSP projects made their first batch of resources available. In this paper we concentrate on one of these projects, CESAR (CEntral and South-east europeAn Resources).

2.1. The CESAR project

Human language technologies crucially depend on language resources and tools that are usable, useful and available. The CESAR project (where 9 partners from 6 countries are involved), in close harmony with the META-NET alliance intends to address this issue by enhancing, upgrading, standardizing and cross-linking a wide variety of language resources and tools and making them available, thus contributing to an open linguistic infrastructure.

The main objective of the project is to make available a comprehensive set of language resources and tools covering Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak. The coverage of these languages brings about an added benefit of the project, anticipating and meeting foreseeable requirements with respect to resources from these languages. Building on a wide range of already existing resources and national or international activities, the project creates, populates and operates a comprehensive language-resource platform enabling and supporting large-scale multi- and cross-lingual products and services. The resources already involved (its number is continuously growing) in the project include interoperable mono- and multilingual speech databases, mono- and bilingual corpora, dictionaries, wordnets and relevant language technology processing tools such as tokenisers, lemmatisers, taggers and parsers.

The main goals of the CESAR project are the following:

- to provide a description of: the national landscape in terms of language use; language-savvy products and services, language technologies and resources; main actors (research, industry, government and society); public policies and programmes; prevailing standards and practices; current level of development, main drivers and roadblocks;
- to contribute to a pan-European digital resources exchange facility by collecting resources and by documenting, linking and upgrading them to agreed standards and guidelines;
- to collaborate with other partner projects,

[1] http://www.meta-share.eu/

in particular concurrent ICT-PSP 6.1 pilot projects and the META-NET network of excellence – and where useful, with other relevant multi-national forums or activities, such as FlaReNET[2] and CLARIN[3] – to ensure consistent approaches, practices and standards facilitating a wider accessibility of, easier access to and reuse of quality language resources and tools;

- to help build and operate broad, non-commercial, community-driven, interconnected repositories that can be used by language researchers, developers and professionals;
- to mobilise national and regional stakeholders, public bodies and funding agencies by raising awareness, organizing meetings and other focused events;
- to reinvigorate cooperation between key technology partners in the region, building on previous collaboration in TELRI[4], MULTEXT-East [5] and other projects;
- to bridge the technological gap between this region and the other parts of Europe by filling obvious and important blind spots in language resources and tools infrastructure.

The main pillars of CESAR activity are seen in enhancement of resources and tools (in size, coverage, precision, recall, accuracy), in adaptation of resources and tools to become compliant with the agreed standards for interoperability, in the upgrade of resources and tools by combining them with other resources and tools in order to achieve the foreseen level of interoperability and in adapting user-interfaces to fulfil user requirements. A special effort is taken to achieve a common standard of involved resources and tools in order to enhance and facilitate the foreseen interoperability between them, as well as to evaluate their license schemes and IPR issues.

The CESAR project aims to stimulate ICT-based cross-lingual communication, collaboration and participation and thereby contribute to the creation of a pan-European digital single market by stimulating ICT-based cross-lingual communication, collaboration and participation. Key resources covered by the CESAR project will be linked and made interoperable using the facilities of the META-SHARE repository. The target user community of the resources practically embraces all stakeholders at the modern digital market: everyday end-users, professional end-users (business, administration, media, education, libraries, etc.), as well as expertise holders (researchers, industrialists, policy makers, etc.) Its concern is a careful investigation of the needs of various types of users – from individual users to large multinational organisations – from the perspective of the current status as well as from the near future prospects.

The project is also contributing valuable resources to META-SHARE, which eventually will become an important component of a language technology marketplace for Human Language Technology (HLT) researchers and developers, language professionals (translators, interpreters, content and software localisation experts, etc.), as well as for industrial players, especially small and medium enterprises, catering for the full development cycle of HLT, from research through to innovative products and services.

The cooperative work on covered resources and tools will be done (and submitted) in three phases. The result of the actions (such as the above mentioned upgrade, standardization, harmonization, linking, cleaning of IPR issues) had been already achieved on the so called “first batch” of

[2] Fostering Language Resources Network (www.flarenet.eu/)

[3] Common Language Resources and Technology Infrastructure (www.clarin.eu/external/)

[4] Trans-European Language Resources Infrastructure (//telri.nytud.hu/)

[5] Multilingual Text Tools and Corpora for Central and Eastern European Languages (nl.ijs.si/)

resources, which was published in early December 2011 (the publication of the "second batch" – actions and metadata descriptions – is scheduled at the end of August 2012 and the publication of the "third batch" at the end of February 2013). The first batch covers 33 resources for six languages and contains 20 corpora (19 written and 1 spoken), 2 dictionaries, 4 wordnets, 1 lexicon and 6 speech databases. This batch was contributed by the project partners, but in the following two batches CESAR plans to involve resources from other, nationally relevant centres to increase the number of enhanced resources.

2.2. Methodology and criteria for the selection of resources

The first step was to develop a methodology by which the identified language resources might be evaluated. A query was distributed among the partners to solicit suggestions on how to approach the evaluation procedure. It was confirmed that no single current methodology can be accepted as a standard. Instead, the consortium developed a list of four general indicators that were considered representative and indicative for the selection of language resources. The indicators determine the general requirements to which the selection should be subjected. Different sets of specific criteria have been defined for each indicator. The indicators are described in the following subsections.

2.2.1. General evaluation of resources

In this indicator, the process of enhancing the resources and tools is carried out in three flows: resource upgrade, extension, and cross-lingual alignment. Among these indicators, further classification is made with respect to the following criteria:

For upgraded resources:

- All selected resources are **state-of-the-art** representatives of their type for a given language (yes, no)
- **Equally valuable representatives** are all **included** in the selection (yes, no)
- Current status of resources have **superior quality** at least on regional level without the need of excessive further development (yes, no)
- Licensing issues allow **free** processing and **access** to resources and resource-related materials or the consortium succeeds in reaching an agreement with respective copyright holders (yes, no)

For extended/linked resources:

- The **extension** of resources **provides considerable value** to the community, at least on regional level (yes, no)
- The emphasis is on **providing building blocks to the existing tools rather than major restructuring** (yes, no)
- **Additional resources are integrated** with the existing ones **only if they significantly improve the quality** of resulting resources (yes, no)
- **If more than one representative of a certain resource type** for a language has been selected, **they are** very likely **to be interlinked** to benefit from strong sides of both solutions (yes, no)
- If **less-developed, but still very popular resources** can benefit from the enhancement due to their well-developed equivalent, their enhancement **is also considered** (yes, no)
- **Experience of other consortium members**/other consortia is extensively used in the process of extension of national resources to provide strong foundation for cross-lingual coverage (yes, no)
- **Tools that are language-neutral** or cross-lingual, **are preferred** (yes, no)

For resources aligned across languages:

- **No more than one tool of a certain type** for each language is used (yes, no)
- Whenever applicable, the **largest set of**

languages is selected (yes, no)

- Language Processing Tools in **NooJ** (yes, no)
- **Language-independence** is targeted to a great extent (yes, no)
- The **quality of a result** is of immense concern (yes, no)

The soundness of specification cannot be judged without knowing the broader context of usage, adequacy, and so on, of the language resource. To estimate the quality, quantity and importance, every case will be thoroughly examined, taking into account regional determinants, popularity of the format outside its home institution, etc. This indicator requires a complex assessment of language resources in the context of the whole set of the established criteria. The partners not only appraise whether the selected resources fulfill the established criteria but also provide concrete examples and detailed explanations based on a thorough analysis.

2.2.2. Total Point Value

Following the approach of the EU NEMLAR (Network for Euro-Mediterranean Language Resources) project (concerning a BLARK for Arabic – for more on BLARK see subsection 2.2.5), the notions of availability, quality, quantity and standards are further specified and taken into account in the process of language resource selection. A technique, supplementing the NEMLAR approach, while defining exact measures for quality and quantity aspects and incorporating the standardisation into the quality section, is developed. The evaluation process consists of the following steps: specification of the point value (PV) of every measure for each resource; aggregation of the points into a single value (total point value, TPV); showing the usefulness of the language resources in further processing; selection of these language resources that fulfil predefined conditions. The following PVs have been specified:

i. Availability: Available for whom?, At what price?, How straightforward it is to reuse it (degree of adaptability)?;

ii. Quality: Standard compliance (Is the resource based on a common standard?), Soundness (Internal consistency, i.e., is the resource based on well-defined specifications?), Task-relevance (Is the resource suited for a specific task?), Environment-relevance (Is the resource interoperable with other resources?);

iii. Quantity (resources only).

The lowest possible TPV is 8, the highest – 25. The established criteria for selecting language resources require TPV lower than or equal to the minimum value of 16. The TPV for resources being selected for the project could be calculated before any upgrade work. The process is directly related to lowering TPVs, which can be used as a concrete indicator of project success.

2.2.3. Language White Papers

The META-NET Language White Paper series “Languages in the European Information Society” reports on the state of each European language with respect to Language Technology and explains the most urgent risks and chances.

The Language White Papers provide an overview of the current situation of language technology support. The rating of existing resources and tools is based on educated estimations by several leading experts using the following criteria (each ranging from 0 to 6).

i. Quantity: Does a tool/resource exist for the language at hand? The more tools/resources exist, the higher the rating;

ii. Availability: Are tools/resources accessible, i.e., are they Open Source, freely usable on any platform or only available at a high price or under very restricted conditions?;

iii. Quality: How well are the respective performance criteria of tools and quality indicators of resources met by the best available tools, applications or resources?;

iv. Coverage: To which degree do the best tools meet the respective coverage criteria? To which degree are resources representative of the target language or sublanguages?;

v. Maturity: Can the tool/resource be considered mature, stable, ready for the market? Can the best available tools/resources be used out-of-the-box or do they have to be adapted?;

vi. Sustainability: How well can the tool/resource be maintained/integrated into current IT systems?;

vii. Adaptability: How well can the best tools or resources be adapted/extended to new tasks/domains/genres/text types/use cases, etc.?

The benefits offered by Language Technology differ from language to language depending on factors such as the complexity of the respective language, the size of its community, and the existence of active research centres in the area. For the resources to be upgraded, quality and maturity of data are the most important factors when no additional processing, apart from automatic conversion, is planned. For resources to be multilingually aligned, all factors are crucial and the resource should represent the best available result. On the other hand, the resources with lowest scores are equally well suited for further processing since low numbers reflect the need of improvement in particular language area.

The situation in covered languages was described and compiled by the CESAR partners – Bulgarian (Blagoeva et al., 2011), Croatian (Tadić et al., 2011), Hungarian (Simon & Lendvai, 2011), Polish (Miłkowski, 2011), Serbian (Vitas et al., 2011), and Slovak (Šimková et al., 2011).

2.2.4. Proportion between the selected resources developed inside and outside the consortium

The resources can be classified as being developed inside consortium, outside consortium or both. This information provides supplementary evidence that can be compared with the gaps – thus, some efforts might be concentrated for further identification of language resources outside the consortium.

2.2.5. BLARK and conclusion

The BLARK (Basic Language Resources Kit) concept was defined by a joint initiative between ELSNET [6] and ELRA[7] . BLARK is defined *as the minimal set of resources that is necessary to do any precompetitive research and education at all*. BLARK includes many different resources, such as (mono- and multilingual) written and spoken language corpora, mono- and bilingual dictionaries, terminology collections and grammars, taggers, morphological analysers, parsers, speech analysers and recognisers, etc. ELDA [8] (Evaluations and Language resources Distribution Agency) elaborated a report defining a (minimal) set of language resources to be made available for as many languages as possible.

BLARK can provide guidelines for prioritising the initial selection and making it independent of local preferences. There is information provided on coverage of language resources and tools for 13 EU languages (both major and lower-density languages) based on the collective experience and expertise gained by the members of the language community. So far, none of the languages of CESAR project are provided with the official BLARK information. On the other hand, Language White Papers classification and Total Point Value calculations for tools and resources for the individual languages in the project give extensive set of evidence for and against the selection of a particular resource. For this

[6] European Network of Excellence in Language and Speech (www.elsnet.org)

[7] European Language Resources Association (www.elra.org)

[8] Evaluations and Language resources Distribution Agency (www.elda.org)

reason, despite our initial intention, BLARK was excluded from the evaluation indicators selected for the CESAR project.

To conclude, the combination of four indicators (each of them specified according to different sets of criteria) are used in the process of selection of CESAR language resources. The first indicator is general, thus assessing the indicator according to general yes/no criteria. All evaluated resources for a given language are listed within the criterion “All selected resources are state-of-the-art representatives of their type”. The next two indicators – Total Point Value and Language White Papers – are based on a numerical assessment of the resources according to previously established qualitative and quantitative criteria and conventions for their measurement. The preferable source of data for our analysis are the tables for individual languages produced by the marks given for each of predefined categories in the Language White Papers. The fourth indicator is complementary – it is not of utmost importance for the selection itself but hints where the efforts should be put to fill the gaps in the selection.

2.3. Licensing issues

Licensing principles and considerations were previously laid down by the 4 consortia – including CESAR – who constitute the actual META-SHARE community. The main goals were to promote the open and free of charge use of data or tools wherever possible, in order to ensure unrestricted access to the resources concerned. Following the Creative Commons and Open Data Commons principles, a set of license schemes was developed and code named “META-SHARE Commons”. Beside these templates, the original Creative Commons (CC) and GPL licenses are also used for some resources. All open licenses allow for further redistribution of the resources or tools. Licenses may have the following standardized and well defined legal attributions (the acronym in parenthesis is popularized by Creative Commons):

- a clause requiring the attribution of the original licensor (BY)
- a clause requiring non-commercial use of the resource/tool (NC)
- a clause requiring the redeposition of all derivative works or products with the same license conditions (called share alike in Creative Commons, SA)
- a clause prohibiting the creation of derivative works (ND)

These additional clauses can be combined, except for SA and ND which are mutually exclusive.

Of course, relicensing an existing resource or tool is often not feasible, especially if there are too many copyright holders involved or there are other restrictions upon the data. Specifically, OpenSource/OpenContent or free licenses tend to be not applicable due to existing legal (mostly copyright), strategic or other considerations. Therefore another license set was developed providing standardized license templates which prohibit any kind of redistribution without the explicit permission of the original licensor. These restrictive licenses can be either commercial or non-commercial, for-a-fee or free of charge or might even have a “No Derivatives” additional restriction.

The CESAR consortium applies the most appropriate license scheme for a given resource out of the set of templates presented so far. Model licenses have been checked by the consortium with respect to regulations and practices at the national level, taking account of possibly different regimes due to ownership (private vs. public sector), type (data vs. software), or pre-existing arrangements with the owners of the original content from which the resource was derived. Resources resulting from the project should be made compliant with the legal principles and provisions established by META-SHARE, as completed/amended by the consortium and accepted by the respective right holders of a given resource or tool.

Optionally, some of the resources/tools are multiply licensed. This concerns mostly Open-Source/OpenContent resources, and typically the set of licenses are GPL, CC-BY-SA and GFDL.

3. Activities and goals

The first batch of resources was due on the 1st December 2011. CESAR consortium partners released 52 language resources or tools by this date, available at the CESAR META-SHARE repository (and also in META-SHARE central repositories). Later the repositories are planned to be synchronized, currently this is not the case yet. As the resources are numerous and are presented below – together with resources and tools scheduled for subsequent 2nd and 3rd batches – we do not provide detailed statistics here due to space reasons.[9]

3.1. Bulgaria

Language resources and tools are being developed at many research centres in Bulgaria – University of Sofia (for example, speech corpora), Plovdiv University (for example, electronic dictionaries), New Bulgarian University (translation memory resources), South-West University (parallel corpora) and others. Some of the most prominent among them are research institutes at the Bulgarian Academy of Sciences – Institute for Mathematics and Informatics, Institute for Communication and Information Technologies and Institute for Bulgarian Language. Several important language resources and tools are being developed at the Institute for Bulgarian language and reported in the scope of the CESAR project. We are going to mention here some of the most prominent language resources, the whole list can be seen at the CESAR and META-NET web sites.

The Bulgarian National Corpus (BulNC) is a publicly available constantly enlarged corpus focused on Bulgarian (currently comprising 469.5 million tokens). It is designed as a uniform framework for texts of different modality (written – spoken), period (synchronic – diachronic), and number of languages (monolingual – parallel where one of the counterparts is Bulgarian).

Thus any language X from the Bulgarian-X Parallel Corpus is equally treated (as a part of the uniform framework of the BulNC) with respect to the text type diversity and balance, metadata description scheme, preprocessing and annotation, data storage format and search engine queries. At present there are texts in 33 languages included in the parallel corpus. However, the languages are not equally represented: the largest parallel corpus is the Bulgarian-English (81.4 million tokens for Bulgarian), there are 21 parallel corpora of size in the range 30–52 million tokens, further in the range 1–10 million tokens, and the rest are below 1 million,

Two manually annotated corpora are also part of the Bulgarian national corpus – BulPoSCor – the POS Tagged Bulgarian Corpus comprising more than 150 thousand words, and BulSemCor – the Sense Tagged Bulgarian corpus containing approximately 100 thousand words. Each lexical item (simple or compound word) which occurs in the particular context in BulSemCor is assigned manually the unique semantic or grammatical meaning from the Bulgarian wordnet. The BulNC search system [10] is designed to support monolingual Bulgarian corpus and parallel corpora in a uniform way. For a given query the system retrieves matches in all documents irrespective of the language. The query language supports terms "word, grammatical feature and relation" (queries for word forms, synonyms,

Please visit the META-SHARE repository of CESAR at http://nlp.ipipan.waw.pl/metashare and see further sections of this document for more information.

[10] http://search.dcl.bas.bg

hypernyms, and similar adjectives are allowed). A web service providing data statistics for collocations is also available.

The Bulgarian wordnet (BulNet) is one of the most complete and consistent lexical resources for Bulgarian (in comparison, the literals in the Bulgarian wordnet are much greater in number than the word list in a standard spelling dictionary). The synonym sets from BulNet are connected by means of inter-language equivalence relations with the Princeton WordNet 3.0, thus it is a part of the multilingual lexical-semantic network, the so called global wordnet. The Bulgarian wordnet is approximately one quarter the size of the English wordnet and is one of the biggest in Europe. The BulNet is distributed by ELDA.

Large Bulgarian morphological dictionaries developed by a number of centres have existed for a long time. They allow for the automatic analysis and synthesising of word forms and thus provide the ability to construct a paradigm (all possible forms) of a given word, the recognition of a given form as a part of a paradigm and to ascribe the grammatical features. The reasons for the selection of the Bulgarian morphological dictionary are that it is at the disposal of the consortium, it is based on the last edition of the Bulgarian orthography dictionary and it is used for the development of a Bulgarian spell checker – one of the targets for a distribution within the CESAR.

3.2. Croatia

In the CESAR project a single Croatian partner is Faculty of Humanities and Social Sciences, University of Zagreb. Its Institute and Department of Linguistics has been involved in producing computer corpora since 1967 and it is considered a leading national institution for corpus and computational linguistics in Croatia. The resources and tools in the first batch of CESAR resources and tools for the Croatian language consist of the following:

Resources:

- *Croatian National Corpus* (HNK) is a representative corpus of contemporary Croatian standard language written texts published since 1990. The corpus is automatically lemmatised and morpho-syntactical description (MSD) tagged using the CroTag hybrid tagging and lemmatising system. 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 the Bonito client is available without any restrictions, while the original texts cannot be distributed for copyright reasons. The Bonito client provides opportunities for complex queries due to the 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. These data are freely downloadable and usable for any further processing.
- *Croatian Morphological Lexicon* (HML) is an inflectional lexicon generated automatically by Croatian Inflectional Generator from ca 110,000 lemmas yielding over 4,000,000 word forms. It has been a result of the work on the basis of the theoretical background published in 1992 (see Tadić 1994). The initial set of lemmas was collected from several existing Croatian mono- and bi-lingual dictionaries, while additional entries were collected via corpus or by means of automatic enlargement of the initial list of lemmas (see Bekavac, Šojat 2005, and Oliver, Tadić 2004). The automatically generated output was corrected for known systemic errors, encoded in UTF-8 and stored in the MULTEXT-East Lexica format, i.e.

lemma[TAB]word-form[TAB]MSD. The MSD-tagset is conformant with the MULTEXT-East v4.0 recommendations for the 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 a PHP script call.

- *Croatian-English Parallel Corpus* (Hr-En p-corp) is a parallel unidirectional (Croatian to English) corpus of contemporary Croatian standard language collected from articles appearing in the Croatia Weekly newspaper, published from 1998 to 2000. The corpus samples were obtained in the digital form entirely, have been converted to XML, aligned using Vanilla Aligner, manually checked and stored in the TMX format. The corpus is available for downloading through META-SHARE distributional platform.
- *Croatian Valency Lexicon of Verbs*, Version 2.0008 (CROVALLEX 2.0008) is an attempt at a formal description of valency frames for Croatian verbs. CROVALLEX 2.0008 was developed as the part of the PhD thesis (Mikelić Preradović, 2008). 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. They were selected from the Croatian frequency dictionary, according to their absolute frequency above 10.

Tools:

- *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 a 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 the contemporary standard Croatian language and returns a zip file with results of processing the uploaded file. At the moment, the limitation of file size is 50,000 tokens. The processing gives the complete analysis for each token, i.e. the line in the verticalised corpus regarding the lemma, POS and MSD. The web interface allows the 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 MULTEXT-East v4.0 specifications for Croatian. 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.

3.3. Hungary

Hungary is represented in the project by two partners, the Research Institute for Linguistics of the Hungarian Academy of Sciences (RIL-HAS) – being also the coordinator of CESAR – and the Department of Telecommunications and Media Informatics at the Budapest University of Technology and Economics (BME-TMIT). The profiles of the two institutes differ, thus the conveyed resources and tools are rather in a complementary composition.

The Research Institute for Linguistics of the Hungarian Academy of Sciences is the leading institute dedicated to Hungarian linguistics (general, theoretical and applied linguistics, Uralic linguistics, and phonetics, as well as the preparation of a comprehensive dictionary of the Hungarian language). It was one of the first centres in Hungary to include large scale computational linguistic work in its agenda.

The Department of Language Technology was created in 1997 (originally as Department of Corpus Linguistics) by its current head Tamás Váradi. The department has been regularly involved in language technology projects ranging from corpus linguistics, shallow parsing, ontology development, machine translation and language resources building. It has accumulated significant research experience and has made remarkable achievements, especially in the development of linguistic resources. It has participated in several successful international projects which were aiming, on the one hand, to adopt certain processes developed for western European languages and now considered part of the standard for the analysis of Hungarian (MULTEXT-East, Gramlex [11]) and, on the other hand, to develop new standards of creating linguistic resources (electronic dictionary databases, CONCEDE[12], CLARIN) as well as machine translation [13]. The researchers at the department have acquired significant knowledge about computerized language processing systems and technologies developed or applied in these projects, and have played an active role in adapting these to the needs of Hungarian.

The seven resources and three tools offered in the first batch are mostly available, but not for the wide public, rather for academic non-commercial use.

Resources:

- *Szeged Corpus* – A morpho-syntactically annotated and manually disambiguated corpus of 1.2 million words (a database of six different topics, approximately 200 thousand words each). The 1.2 million words cover 155,500 different word forms, and also contain further 250 thousand punctuation marks. Corpus files are available in the XML format, their inner structure is described by the TEIx-Lite DTD (Document Type Definition) scheme.
- *Szeged Treebank* – A manually checked treebank of 1.2 million words. The determination of marked syntagmas and their relationship helps further linguistic processing, among others the semantic analysis of texts. There was made an intensive mark-up of syntactic structures on 82 000 sentences (1.2 million word entries + 250 thousand punctuation marks) of the Szeged Corpus 2.0 file. Treebank files are stored in the XML format, their inner structure is described by TEI P4 DTD scheme.
- *Szeged Named Entity Recognition Corpus* – A manually annotated part of the Szeged Treebank, consisting of short business news. The used NER categories are (based on the CoNLL system) the following: PERSON, ORGANISATION, LOCATION and OTHER.
- *Hungarian WordNet* – The Hungarian WordNet is a multilingual ontology, meaning that most of its synsets were mapped to equivalent concepts in English (Princeton) WordNet v.2.0. The ontology is also linked to entries of a Hungarian

[11] Lexiques grammaticaux et morphologiques (www-igm.univ-mlv.fr/~laporte/Copernicus/)

[12] Consortium for Central European Dictionary Encoding (www.itri.brighton.ac.uk/projects/concede)

[13] Internet Translators for all European Languages (iTranslate4.eu)

Monolingual explanatory dictionary and to the entries of the Hungarian verb valency frame lexicon.

- *Hungarian Webcorpus* – A corpus of over 1.48 billion words (from which 589 million words are morphologically filtered), this is by far the largest Hungarian language corpus, and it is available in its entirety under a permissive Open Content license. The Hungarian Webcorpus was created as part of the WordSword project at the Media Research and Education Centre and consists of 18 million pages downloaded from .hu domain.
- *Hunglish Corpus* – The Hunglish Parallel Corpus is a free sentence-aligned Hungarian-English parallel corpus of about 2 million sentences. The corpus may be searched through a web-based sentence search service. This service has more than 200,000 visits per month.
- *morphdb.hu* – Hungarian lexical database and morphological grammar. The morphdb.hu is described in the formal representation form of hunlex, an offline resource compiler which offers a linguistically motivated morphological description language and allows for principled, flexible maintenance and extension of resources. The morphdb.hu thus provides – with the help of h3unlex and hunmorph – primary language resources for spell-checking, stemming, morphological analysis and numerous other annotation tasks.

Tools:

- *hunalign* – A powerful free sentence level aligner tool for building parallel corpora. The tool was developed under the Hunglish project to build the Hunglish Corpus. The hunalign aligns bilingual texts at the sentence level. Its input is a tokenized and sentence-segmented text in two languages. In the simplest case, its output is a sequence of bilingual sentence pairs (bisentences). In the presence of a dictionary, it uses it, combining this information with Gale-Church sentence-length information. In the absence of a dictionary, it first falls back to sentence-length information, and builds an automatic dictionary based on this alignment. Then it realigns the text in a second pass, using the automatic dictionary. Like most aligners, hunalign does not deal with changes of sentence order: it is unable to come up with crossing alignments, i.e., segments A and B in one language corresponding to segments B' and A' in the other language.
- *hunmorph* – An open source tool and programming library for spell-checking, stemming and morphological analysis developed mainly for agglutinative languages. The hunmorph is based on an extention of the codebase of MySpell, a reimplementation of the well-known Ispell spellchecker, yielding a generic word analysis library. At this point, the development of the library has forked. Now the extended MySpell, called HunSpell, is part of the LibreOffice (and was also of OpenOffice.org) multilingual office suite. The hunmorph is the program tuned to morphological analysis. The hunmorph framework is built from three components: a) the ocamorph runtime analyzer is a language independent affix stripping implementation, b) morphdb.hu is a lexical database and morphological grammar, which can be used by ocamorph (details can be found at http://mokk.bme.hu/resources/morphdb.hu), c) the hunlex is an off-line resource management component, which complements the efficiency of our runtime layer with a high-level description language and a configurable

precompiler.

- *huntoken* – A rule based tokenizer and sentence boundary detector for Hungarian (and English) texts. Its input is a plain text file in the ISO Latin-1 or Latin-2 character encoding and the output is a tokenized XML file. It determines the correct word and sentence boundaries with 98% precision. It can be used under Unix, Linux, Mac OSX and Windows, as well.

BME-TMIT is a key player in speech technology research and applications in Hungary, its speech technology group was founded in 1969 by Prof. Géza Gordos. The main profile of the group is speech technology research in speech recognition, speech synthesis and speech databases. Several corpora and tools were developed and created at TMIT, including general and specialized corpora for speech synthesis and speech recognition or speech technology research and several tools for audio and text processing (synthesis and recognition engines, phonetic transcribers, forced aligners, prosodic segmenter, hearing training, etc.)

A set of these resources and tools is offered to be shared via META-SHARE. Our general licensing guidelines are to support open and possibly free access of resources or tools for non-commercial and academic exploitation. Nevertheless, several resources are or will be available for commercial purposes, too. Our contribution includes:

- **Speech corpora for research related to speech technology and speech communication (also including non-verbal communication):**

BABEL – Hungarian Clear Speech corpus is an EUROM1 compatible database containing records from 60 speakers distributed in 3 speakers set.

Emotion Database – Spoken corpus holding emotionally rich utterances, labeling is done for emotions (8 basic emotions are labeled).

Sound Gesture Database – a lexicon of sound gestures consisting of 770 tokens.

Medical Database – The Medical database is a speech corpus holding utterances from persons suffering from different speech problems of organic origin.

Formant database from spoken words – speech corpus for Hungarian, it is manually annotated. The formant database is constructed from 3000 Hungarian spoken word items, for research and education. It is a reference database for Hungarian formant data.

● **General speech corpora for speech recognition or speech synthesis:**

MRBA – The Hungarian Reference Speech Corpus contains continuous read speech. The database contains utterances read by 332 different speakers. The utterances were recorded in acoustically different locations.

MTBA – The Hungarian Telephone Speech Corpus is a PSTN and mobile telephone voice Hungarian speech database. The database contains records based on the definition in SpeechDatE for the dialectical, age and sex balance and vocabulary. What is important and different from the SpeechDatE database is that the phonetically rich sentences and words have been segmented and labelled at the phoneme level.

MTÜBA – Hungarian Telephone Client Speech Corpus contains telephone calls recorded at the call centre of a service provider company. The corpus consists of dialogues between the operator and the client. The orthographic transcription of the speech utterances is provided, clauses are segmented (automatically followed by hand-correction).

Hungarian MALACH – Hungarian Speech Database of Holocaust Survivors' Testimonies, i.e., World War II histories of elderly people, typically holocaust survivors.

Hungarian Parliamentary Speeches – publicly available corpus with approximate transcriptions of Parliamentary Speeches.

Di-phone database for TTS – a spoken speech corpus for Hungarian, German, Spanish, manually annotated and consisting of 5500 records. Diphone logatom-based sound recordings for Text-to-Speech (TTS) conversion. Voices for Hungarian: male and female, for German: male, for Spanish: male.

Word level speech database – a speech corpus for Hungarian, manually annotated, contains 2 hours of speech. It is a read word list for the presentation of the segmental level structure of Hungarian speech (the CV,VC,VV,VVV,CC,CCC,CCCC clusters). Voices: male and female.

Read speech database for TTS – a speech corpus for Hungarian, semi-automatically annotated, contains 20 hours of speech. Contains read sentences for unit selection TTS and for Hidden Markov Model (HMM) based TTS systems. Voices: male and female.

Named entity lexical database – a lexicon transcribed for TTS, contains collection and transcription into spoken form of names (person, legal entities, locations) for Hungarian name and address reader concept-to-speech.

Spoken number database for TTS – a speech lexicon for Hungarian, German, English (3 hours). It contains number elements for high quality number, time, date reading in (e.g. financial) information systems. Voices in Hungarian: male, female, in German: male, in English: male.

Hungarian Pronunciation Dictionary – Pronunciation vocabulary for Hungarian word forms. This is the only reference pronunciation database for Hungarian.

● **Multimodal corpora:**

Broadcast News Database – This corpus was constructed for 10 European languages within the Broadcast News Interest Group of the COST278 action (Žibert et al. 2005). The Hungarian material consists of 3 hours and 30 minutes of recordings, transcribed and annotated, using the conventions of NIST (National Institute of Standards and Technology, USA).

BLD corpus – The Broadcast Lectures Database contains recorded Broadcast Video Lectures from wide scientific topics for the public.

3.4. Poland

Polish resources and tools are being contributed to CESAR by two Polish partners: Institute of Computer Science of the Polish Academy of Sciences and University of Łódź.

The Institute of Computer Science, Polish Academy of Sciences (IPIPAN – http://www.ipipan.eu/) is a leading national center of fundamental research in Computer Science as well as applied research in the areas of Artificial Intelligence and Information Systems. Among the tools and solutions developed within the Linguistic Engineering Group [14] of IPIPAN are taggers, shallow and deep parsers, as well as various machine learning and rule-based information extraction tools. IPIPAN has also developed large, linguistically annotated corpora which have found numerous applications in Information Extraction and Text Mining both in general and special domains.

University of Łódź is another major research and education center in Poland. The PELCRA[15] research group based at the university's department of linguistics is a long-standing player on the Polish language resources scene. Over the recent years, its activities have been geared towards the collection of corpus data, including multimodal spoken data and development of language tools and services with applications in research and technology. The group's members have confirmed experience in developing lan-

[14] http://zil.ipipan.waw.pl/

[15] http://pelcra.ia.uni.lodz.pl/

guage processing systems for both general and special domain Polish and English texts.

The first batch contained 10 Polish resources and tools:

- *PoliMorf Inflectional Dictionary* (preliminary version 0.5) – the new morphological dictionary for Polish resulting from the standardization, merger and automated correction of two most important morphological dictionaries of Polish – Morfeusz SGJP and Morfologik, containing 6.8 million inflected word forms with their morphological descriptions. Next versions of the resource will contain manually corrected data. The first of the merged resources, Morfeusz SGJP, is based on SGJP – Grammatical Dictionary of Polish – which is the result of long-standing work of an informal group lead by Prof. Zygmunt Saloni. The work started in the 1980s by digitising the list of headwords of the 11-volume Doroszewski's dictionary of Polish (1958–1969). The grammatical description in SGJP is based on new concepts proposed in the 2nd half of the 20th century with many detailed solutions proposed by the members of the team (Tokarski, Gruszczyński, Saloni). PoliMorf uses data from the second edition of SGJP. 244,341 lexemes correspond to 4,223,981 word forms (counting syncretic forms of the same lexeme as one unit). Inflection in SGJP is represented with inflectional patterns, which describe forms in terms of a stem common to all forms and endings differentiating the forms.

  The second merged resource, Morfologik, is another open-source morphological dictionary of Polish. It contains 216,992 lexemes and 3,475,809 word forms. The dictionary was created by enriching the Polish ispell/hunspell dictionary with morphological information. Unfortunately, the original source dictionary did not contain sufficient structure to allow reliable detection of some information, such as the exact subgender of the masculine for substantives. This information was added manually and using heuristic methods. The tagset of the dictionary is inspired by the IPI PAN Tagset. However, Morfologik diverges from that tagset and from Morfeusz, as it never splits orthographic ("space-to-space") words into smaller dictionary words (i.e. so-called agglutination is not considered). Moreover, due to the lack of information in the ispell dictionary, some forms are not completely annotated, and are marked as irregular. There is, however, some additional markup added to reflexive verbs, which is not present in the original IPI PAN Tagset. This was introduced for the purposes of the grammar checker LanguageTool that used the dictionary extensively.
- *1 million subcorpus of National Corpus of Polish* – manually annotated sample of the 1.5-billion National Corpus of Polish (Narodowy Korpus Języka Polskiego, NKJP) [16] – a shared initiative of four institutions: Institute of Computer Science at the Polish Academy of Sciences (coordinator), Institute of Polish Language at the Polish Academy of Sciences, Polish Scientific Publishers PWN, and the Department of Computational and Corpus Linguistics at the University of Łódź. It has been registered as a research-development project of the Ministry of Science and Higher Education. The list of sources for the corpus contains classic literature, daily newspapers, specialist periodicals

[16] http://nkjp.pl/

and journals, transcripts of conversations, and a variety of short-lived and internet texts. The resources represent wide diversity with respect to the subject and genre. The spoken part covers both male and female speakers, in various age groups, coming from various regions in Poland.

- *Polish Sejm Corpus* (version 1.0) – a collection of annotated utterances of Polish Sejm members from terms of office 1–6 (years 1991–2011). Corpus data, available in an NKJP variation of the TEI P5 format, [17] contain information about text segmentation (paragraphs, sentences, tokens), disambiguated morphosyntactic description (lemma, POS tag, MSD tag), syntactic description (syntactic words and groups) and named entities (person names, locations, organization). The data is a valuable source of linguistic information, being a large (114 million segments) collection of quasi-spoken content and making the basis of the audio/video recording of sessions, started in 2011 and planned to be consecutively appended to the corpus. Next versions of the corpus will be extended with related content – parliamentary questions and Sejm committee meeting transcripts.
- *Polish WordNet* (Słowosieć, version 1.5) – a network of lexical-semantic relations, an electronic thesaurus developed at the Wrocław University of Technology, with a structure modelled on that of the Princeton WordNet and those constructed in the EuroWordNet project. Polish WordNet describes meaning of a lexical unit by placing it within a network of semantic relations, such as hypernymy, meronymy, antonymy, etc. To reduce the cost of the project, Polish WordNet has been built semi-automatically. Lexical relations were automatically recognized in large corpora of Polish and suggested to lexicographers via a graphical interface. Nowadays Polish WordNet is one of the biggest wordnets in the world; it comprises 103,000 lexical units in 74,000 synsets.
- *Polish Named Entity Recognition Tool* (NERF, version 0.2) – a statistical tool for Named Entity Recognition (NER) based on the Conditional Random Fields (CRF) modelling method. The tool has been constructed as a part of the National Corpus of Polish project. It has been adapted to recognize tree-like structures of NEs (i.e., with recursively embedded NEs) using the Joined Label Tagging (JLT) method. The JLT method is a simple method of encoding NE structures as a sequence of labels. With this method various additional information about NEs of categorical nature – type, subtype, type of derivation – can be encoded on the level of labels and subsequently recognized using the resultant CRF model. The tool can be configured to use various types of observations during the training and recognition process, for example: lexical information from textual level, or grammatical information from morphosyntactic level.
- *Polish Named Entity Resources* (preliminary edition) – a set of named entity-related resources obtained from existing sources and supplemented with additional language-specific resources acquired from the Web. Whenever appropriate, inflected forms were generated using Morfeusz SGJP generator. The Polish Named Entity Resource data was used in the process of named entity annotation of the National Corpus of Polish (NKJP). The resource data is described with complex

[17] http://nlp.ipipan.waw.pl/TEI4NKJP/

hierarchy of names: forenames and surnames, city, country, mountain, region and river names, institution names, relational adjectives and inhabitant names stemming from country names, named entity triggers (months, days, positions, etc.). Currently it contains 153 thousand entries.

- *LUNA.PL corpus* contains 500 human-human spoken dialogues in Polish (13,000 utterances). The corpus is annotated on several levels, from transcription of dialogues and their morphosyntactic analysis, to semantic annotation on concepts, predicates and anaphora. Annotation on the morphosyntactic and semantic levels was done automatically and then manually corrected. At the conceptual level, the annotation scheme comprises about 200 concepts from an ontology designed specifically for the project. The set of frames for predicate level annotation was defined as a FrameNet-like resource.
- *LUNA-WOZ.PL corpus* contains 69 human-computer spoken dialogues in Polish (5.5 thousand utterances). The corpus is annotated on several levels, from transcription of dialogues and their morphosyntactic analysis, to the semantic annotation of concepts.
- The largest collection of transcriptions of naturally occurring conversational Polish has been compiled by the PELCRA team at the University of Łódź since 2000, initially as part of the PELCRA reference Corpus and later within the National Corpus of Polish. In total, the corpus contains almost 2 million words of transcriptions of conversations recorded in an informal setting, often without some of the speakers knowing they were being taped (although they had been informed about and agreed to the possibility of being recorded and later granted their permission to transcribe the recordings). So far this data has been only available through online search interfaces, but within CESAR a subset of this data totaling 1.8 million words has been made available in the TEI P5 format following some privacy considerations. Furthermore, a selection of the transcriptions (at least 200 000 words) are being time-aligned with the original recordings at the level of utterances and made available under a CC-like license through the META-SHARE repository as a time-annotated multimedia corpus of conversational spoken Polish. Such resources are undoubtedly a very precious source of linguistic data with possible applications in modeling conversational Polish and special settings speech recognition.
- Although a number of freely available public domain and open license parallel resources exist for Polish, they generally suffer from problems which seriously affect their usability and interoperability. One of the responsibilities of the University of Łódź within the Polish branch of the CESAR project is to compile a pool of Polish parallel corpora in a number of standard formats which will make them readily available for NLP and computer-assisted translation purposes. So far four sets of parallel corpora have been released in the TEI 5 and XLiFF formats under the Creative Commons license. Some of those corpora have so far been available, but lacked bibliographic annotation (Acquis Communautaire), others are new resources compiled from freely available resources (CORDIS, RAPID), but there are also parallel texts acquired directly from publishers and manually aligned at the sentence level (the Academia magazine). The latter resource required special

copyright arrangements with the publisher and it is also a good illustration of CESAR's impact in terms of clearing the intellectual property rights status of LRTs to make them available for use.

Apart from LRT delivery, the Polish CESAR partners succeeded in increasing the awareness of existing LRTs and reinforcing relations between the key players in Polish natural language processing with establishing a new Web portal "Computational Linguistics in Poland" (CLIP, http://clip.ipipan.waw.pl) in April 2011. The site contains exhaustive information about LRTs, research centres, projects and linguistic engineering courses related to Polish. Furthermore, it intends to bring language-related initiatives, institutions and people from research, government and industry communities together, offering them comprehensive information on available language technology. One of the main design principles of the site was to maintain a wiki-like mode of operation, allowing the authorised representatives of all LRT groups in Poland to edit the content directly. This approach proved very fruitful and several modifications and additions have already been made by external editors. According to our best knowledge the site is currently the largest repository of references to publicly available Polish LRTs.

3.5. Serbia

The Human Language Technology Group hosted by the Department for Computer Sciences at the Faculty of Mathematics, University of Belgrade, was founded more than 30 years ago, in 1978, with the main goal of developing a formal description of Serbian, and producing and exploiting resources and tools for this language. The core of the HLT Group is now composed of researchers from several faculties of the University of Belgrade but it has strong relations with the majority of Serbian institutions involved in language technology.

In the course of its long existence the HLT Group has developed a considerable amount of language resources and tools, among which the most important are:

- Electronic dictionaries, like Serbian WordNet and morphological dictionaries of general lexica and proper names, simple and multi-unit forms;
- Monolingual, bilingual and multilingual text corpora, with general coverage or aiming at specific domains;
- Various tools for developing and/or exploiting these resources, of which the most versatile is *LeXimir*, for development of lexical and textual resources and its web counterpart *VebRanka* for expending search queries.

Some resources of importance for Serbian have also been developed within other parts of the Serbian academic community. Major researchers in the field of Speech Interaction technology are located in the Technological faculty of Novi Sad, and its spin-off, the AlfaNum company. They developed, in addition to speech databases, a lexical database with more than 4,000,000 accentuated word forms for Serbian. Various commercial applications in the fields of Text-to-Speech (TTS) and Automatic Speech Recognition (ASR) have been developed based on these resources. Some very important multimodal resources are developed by the Institute for Balkan Studies of the Serbian Academy of Sciences and Arts in Belgrade. This database contains digitized audio, video, photo and textual ethnographic material collected in Serbia in the scope of the field research.

The role of the HLT Group in the CESAR project is to detect all resources and tools being produced for Serbian, and make them visible not only to the research community but to producers and users of language technology at large. For the first batch of resources delivered to META-SHARE, the HLT Group has decided to prepare its own resources that are not only the most mature but also the most used or asked for by various users. The selected resources are:

- *The Corpus of Contemporary Serbian (SrpKor)* – consists of 4,523 texts with the total size of more than 113 million words. [18] It is lemmatized and PoS tagged using TreeTagger. Texts that entered SrpKor consist of fiction written by Serbian authors in 20th and 21th century, various scientific texts from various domains (both humanities and sciences), legislative texts and general texts (newspapers, journals, magazines, feuilletons). It is available via a web interface since 2003 and it is regularly used my more than 300 users, mostly Slavists, from all over the world.
- *Serbian WordNet (SrpWN)* – represents a hierarchical lexical semantic network, containing synsets with glosses and various semantic relations, such as antonymy, meronymy, causation, category domain, etc. The initial version of the Serbian Wordnet was produced in the scope of the EU-funded Balkanet project.
- *The Serbian Lematized and PoS Annotated Corpus (SrpLemKor)* – consists of a sample of various texts from SrpKor, lemmatized and PoS tagged using TreeTagger. It consists of daily news published in newspaper "Politika", some newspaper feuilletons published in newspapers "Politika" and "Danas", fiction written by Serbian authors in 20th century, and various scientific texts from various domains and legislative texts. Contrary to SrpKor that can be accessed and searched only through the web interface this corpus can be downloaded.
- *French-Serbian Aligned Corpus (SrpFranKor)* – includes French or Serbian source literary and newspaper texts and their translations. The alignment was performed on the sub-sentence level. The corpus contains 25 literary texts and 15 newspaper articles. It can be accessed and searched through the web interface. For more information see (Utvic, 2011) in this issue.
- *Multilingual Edition of Verne's Novel "Around the World in 80 Days" (Verne80days)* – contains 17 editions of Jules Verne's novel "Around the World in 80 Days" – French original and 16 translations. Translations are aligned with French, English or Serbian version, transferred to the TMX format and this version can be downloaded.
- *Organizing digitized material (InfoBeaver)* – an application for collecting and presenting multimedia information. It works with multimedia documents and enables database search using different criteria. The demo-version illustrates its functionalities with some data about CESAR project and its participants.

All resources are offered under CC BY-NC, while the tool is offered under GPL. We are glad to say that some of resources that were made available in the first batch have already been retrieved and accessed via META-SHARE by a party outside Serbia.

3.6. Slovakia

Most of the NLP in Slovakia, especially concerning resources, is concentrated in the Slovak National Corpus department of the Ľ. Štúr Institute of Linguistics. The Institute collected the single most relevant source of data of the modern written Slovak language, the Slovak National Corpus database. Since the establishment in 2003, the policy of the department has always been to release the tools and resources to the professional

[18] korpus.matf.bg.ac.rs

and general public under permissive Open Source and Open Content licences, if possible within existing (mostly copyright related) constrains, and to promote the importance of NLP and digital resources in general and applied linguistics and in other related areas. The policy is continued in the CESAR project – in particular, submitting the resources into the META-SHARE network is deliberately used to clear up the existing licence agreements and to push reluctant copyright holders towards Open Content licensing.

The most visible NLP resource is the Slovak National Corpus – a comprehensive, representative corpus of modern written Slovak (Garabík 2010). The corpus is an ongoing project, at the time of writing containing over 770 million tokens, with automatic morphosyntactical analysis and lemmatisation. The corpus project was in fact the first big, focused project in NLP research in Slovakia and marked the new era of Slovak computational linguistics. This is the reason why many of the resources are somewhat corpus centric – in the previous years, a lot of work had to be done from scratch, since there were few previous usable NLP resources and tools for the Slovak language.

For the META-SHARE, we have selected the resources that are well developed, representative of the language and useful (or even indispensable) for NLP related research and applications. The first and foremost is the Slovak National Corpus, version 5.0. Unfortunately, due to copyright restrictions (the texts in the corpus are covered by existing license agreements with copyright holders, but the agreements cover only their inclusion in the corpus itself, with access only via a search query interface) the corpus cannot be redistributed. Therefore it is accessible only as a pseudocorpus, accessible either via a specialized Tcl/Tk client (Bonito) or via a web interface (Bonito2).

This corpus is complemented by the Corpus of Spoken Slovak (Šimková et al. 2008), version 3.0, containing 300 hours of manually transcribed recordings of standard spoken Slovak, which amounts to 1.6 million tokens. Unlike the written language corpus, this corpus contains data mostly recorded on purpose by the Institute. Since the common narration is not copyrightable (unless it contains an artistic element, e.g. reciting a poem), and all the speakers agreed with the inclusion of their recordings in the database, the corpus is triple-licensed under combination of Affero GPL, CreativeCommons Attribution-ShareAlike and GNU Free Documentation Licence.

Fundamental resource necessary for any language processing of Slovak (a rich morphology language) is a database of word forms, corresponding lemmas and morphosyntactic tags (Garabík 2005). The database contains 77 thousand lemmas, giving 2.5 million word forms.

Parallel corpora are represented by two most developed databases. The first one is the Slovak-Czech parallel corpus, containing 1.5 million aligned sentence pairs. The corpus consists of three types of translations (mostly fiction, but also some non-fiction texts): translations from Czech into Slovak, from Slovak into Czech, and from third language into both Czech and Slovak. The second large parallel corpus is the Slovak-English corpus, containing 700 thousand aligned sentence pairs. The corpus is almost exclusively English language fiction translated into Slovak, but there are some non-fiction texts as well.

Further plans include several other well developed resources (either produced within Ľ. Štúr Institute of Linguistics or elsewhere). These resources are the Slovak Language Treebank, Slovak WordNet, other parallel corpora, dictionary of Slovak collocations and others. The CESAR and META-SHARE is a good opportunity to promote the ideas of open copyright licenses and to clear up older resources and to bring them up to contemporary standards, and also emphasise language technology in Slovakia.

4. Conclusions

One of the central aims of the CESAR project is to build an open language technology infrastructure that will enable the achievements of the respective language technologies of the participating countries to be globally accessible through META-SHARE, the Open Resource Exchange Facility. Language resources and tools in countries participating in the CESAR project have been developed separately, without paying attention to any common infrastructure or interoperability issues. The CESAR project through the META-SHARE repository brought forward the most relevant and most developed data for all the major European languages and attempts to present them through a unified interface and infrastructure. Thus, the CESAR project represents a significant step towards opening up and making accessible in a standardised and fully documented way language resources that would serve to strengthen the European Research Area.

References

Božo Bekavac and Krešimir Šojat. 2005. Lexical acquisition through particular adjectival endings for Croatian. Proceedings of the Workshop on Computational Modeling of Lexical Acquisition, University of Split, Split, 2005.

Broeder, Daan, Thierry Declerck, Erhard W. Hinrichs, Stelios Piperidis, Laurent Romary, Nicoletta Calzolari and Peter Wittenburg. 2008. *Foundation of a Component-based Flexible Registry for Language Resources and Technology*, In Calzolari, N. (Ed.) *Proceedings of the 6th International Conference of Language Resources and Evaluation*, Ed. Nicoletta Calzolari, 1433-1436. European Language Resources Association (ELRA).

Blagoeva, Diana, Svetla Koeva and Vladko Murdarov. 2011. *Languages in the European Information Society – Bulgarian*, META-NET White Paper Series, Berlin.

Garabík, Radovan, Svetla Koeva, Maciej Ogrodniczuk, Marko Tadić, Tamas Váradi and Duško Vitas. 2011. Detecting Gaps in Language Resources and Tools in the Project CESAR, *Human Language Technologies as a Challenge for Computer Science and Linguistics*, Ed. Zygmunt Vetulani, 37-41, , Poznań: Fundacja Uniwersytetu im. A. Mickiewicza.

Garabík, Radovan. 2005. Levenshtein Edit Operations as a Base for a Morphology Analyzer. In: Computer Treatment of Slavic and East European Languages. *Proceedings of the conference Slovko 2005*, Ed. Radovan Garabík. Bratislava: Veda.

Garabík, Radovan. 2010. Slovak National Corpus tools and resources. In: *Proceedings of the 5th Workshop on Intelligent and Knowledge oriented Technologies (WIKT 2010)*. Eds. Laclavík, M., Hluchý, L., 2 – 7, Bratislava.

Mikelić Preradović, Nives. 2010. *Approaches to the Development of the Machine Lexicon for Croatian Language*, PhD Thesis, University of Zagreb, Faculty of Humanities and Social Sciences,.

Miłkowski, M. 2011. *Languages in the European Information Society – Polish*, META-NET White Paper Series, Berlin.

Oliver, Antonije and Marko Tadić. 2004. Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Vol. IV, 1259-1262. Genoa-Paris: European Language Resources Association (ELRA).

Simon, Eszter and Lendvai, P. 2011. *Languages in the European Information Society – Hungarian*, META-NET White Paper Series, Berlin.

Šimková, Mária, Radovan Garabík, Agáta Karčová, Katarína Gajdošová, Michal Laclavík, Jozef Juhár, Karol Furdík, Peter Ďurčo, Helena Ivoríková, Jozef Ivanecký and Július Zimmermann. 2011. *Languages in the European Information Society – Slovak*, META-NET White Paper Series, Berlin.

Šimková, Mária, Radovan Garabík, Agáta Karčová and Katarína Gajdošová: Hovorený korpus slovenčiny. 2008. In: *Čeština v mluveném korpusu*. Praha: Nakladatelství Lidové noviny / Ústav českého národního korpusu, 227–233.

Tadić, Marko, Dunja Brozović-Rončević and Amir Kapetanović. 2011. *Languages in the European Information Society – Croatian*, META-NET White Paper Series, Berlin.

Tadić, Marko. *1994. Računalna obradba morfologije hrvatskoga književnoga jezika*. PhD Thesis, University of Zagreb, Faculty of Humanities and Social Sciences.

Utvić, Miloš. 2011. Annotating the Corpus of Contemporary Serbian, *Infotheca* 12(2), 2011.

Varadi, Tamas. 2011. Introducing the CESAR project. *Infotheca* 12(1), 71-74.

Vitas, Duško., Ljuba Popović, Cvetana Krstev, Mladen Stanojević and Ivan Obradović. 2011. *Languages in the European Information Society – Serbian*, META-NET White Paper Series, Berlin.

Žibert, J., et al. 2005. The COST278 Broadcast News Segmentation and Speaker Clustering Evaluation - Overview, Methodology, Systems, Results. In: *Eurospeech 2005: 9th European Conference on Speech Communication and Technology*. Lisboa, Portugal, 629-632.