CESAR

Central and South-East European Resources
Project no. 271022

Deliverable D2.1
Language WhitePapers

Version No. 1.2
30/07/2011

Document Information

Deliverable number:	D2.1
Deliverable title:	Language WhitePaper
Due date of deliverable:	31 July 2011
Actual submission date of deliverable:	30 July 2011
Main Author(s):	Svetla Koeva (IBL)
Participants:	Tamas Varadi (HASRIL) Tibor Pintar (HASRIL) Radovam Garabik (LSIL) Piotr Rezik (ULODZ) Adam Przepiórkowski (IPIPAN) Marko Tadic (FFZG) Dusko Vitas (UBG)
Internal reviewer:	Tamas Varadi (HASRIL)
Workpackage:	2
Workpackage title:	Analysis and selection of language resources
Workpackage leader:	IBL
Dissemination Level:	Public
Version:	1.2
Keywords:	Language whitepaper

History of Versions

Version	**Date**	**Status**	**Name of the Author (Partner)**	**Contributions**	**Description/ Approval Level**
1.2	29/07/ 2011	PU	Tibor Pinter		editing of the text
1.1	28/07/ 2011	PU	Svetla Koeva		compilation of deliverable

EXECUTIVE SUMMARY

The deliverable describes and summarises the language whitepapers made for six languages on English and local languages. With a special effort describes the Availability of tools and resources as well as a conclusion for each language (situation). The deliverable contains the Language Whitepapers in English and in local languages as an attachment.

Table of Contents

1. Description

The 10th volume of the META-NET Language WhitePaper series "Languages in the European Information Society" reports on the Language Technology (LT) development for the European languages - the Deliverable 2.1 reports on Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak. The Language WhitePaper documents sketch the main findings and challenges facing the LT for Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak, while reviewing the most urgent risks and chances in the field. The volumes focus on the readiness of core technologies for the respective languages.

The first chapter A Risk for Our Languages and a Challenge for Language Technology informs about some of the risks for the European languages after the digital revolution that has changed communication and society. It also lays down a number of challenges facing the implementation of the LT tools and resources, such as language borders, multilingualism, and the slow pace of technological progress. It further explains the LT as a key enabling technology that helps people collaborate, conduct business, share knowledge and participate in society across different languages. The analysis reviews the LT application fields, such as automatic translation, content production, information processing and knowledge management. LT is a tremendous opportunity for the EU citizens to communicate across the language borders on the European common market and global market. The section further explains how computers handle language, taking a brief look at the way humans acquire languages, and then considering how machine translation systems work. The two main types of LT systems – statistical approaches and rule-based systems – acquire language capabilities in a similar manner as humans.

The second chapter Bulgarian/Croatian/Hungarian/Polish/Serbian/Slovak in the European Information Society presents in brief a couple of general facts about the respective language, including features of orthography, morphology (rich inflectional and derivational systems; aspectual verb pairs), syntax (pro-drop, relatively free word order) that may impede the development of the LT tools and resources. Further, it outlines the recent changes in Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak such as internationalization (mostly involving English loan words). The documents discuss the status of Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak as official languages in Bulgaria, Croatia, Hungary, Poland, Serbia and Slovakia and administrative languages in the EU, as well as their inclusion in the respective educational systems and curricula (at elementary, secondary and university level). A further view extends to the Internet resources in Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak (news portals, Wikipedia, web search and online translation services).

The third chapter Language Technology Support for Bulgarian/Croatian/Hungarian/ Polish/Serbian/Slovak addresses the state-of-art of the LT support for Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak. First, it gives a brief account on the LT application architectures that consist of several components to mirror different aspects of language. Second, it covers the situation in the LT research and education. Third, it concludes with an overview of the past and ongoing research programs in universities and institutions. The section ends with an expert estimation of the situation with core LT tools and resources. Spelling checkers and grammar checkers are available for some of the languages, although their precision is still unsatisfactory. A number of applications for web search are also presented reviewing their principles, merits and shortcomings. However, for a more sophisticated request for information, integrating deeper linguistic knowledge is essential.

Consequently, the next generation of search engines will involve much more sophisticated language technology. It may involve subfields such as machine-reading thesauri and ontological language resources, information retrieval, named-entity recognition, among others. Retrieving relevant answers requires an analysis of the sentence on syntactic and semantic level as well as the availability of an index that allows for a fast retrieval of relevant documents. The processed query needs to be matched against a huge amount of unstructured data. Matching a query to documents written in a different language (the so-called cross-lingual information retrieval), is even more demanding.

The analysis proceeds to speech interaction technologies, such as voice user interfaces and text-to-speech applications, for the Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak. In this domain, a genuine market for the linguistic core technologies for syntactic and semantic analysis does in general not exist yet. Another LT application field is machine translation that is particularly challenging for all languages. For Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak machine translation, question answering, information extraction, and summarization (text generation or summary generation) have not been the center of numerous initiatives.

The section also addresses the LT hidden principles. Building language technology applications involves a range of subtasks that do not always surface at the level of interaction with the user but provide significant hidden functionalities.

The WhitePapers shed light on the LT in Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak education including subjects and curricula, mostly at university level. Further, the WhitePapers mention in brief the various programs and initiatives that fund the development of the LT tools and resources for languages in question.

The documents show the results of a survey of the state-of-art of LT tools and resources for Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak in comparison with other languages covered by the META-NET initiative analysing criteria such as quantity, availability, quality, coverage, maturity, sustainability, and adaptability. The results are included into tables of tools and resources for respective languages.

The fourth chapter About META-NET contains an overview of the META-NET tasks, initiatives, and member organisations.

2. Availability of tools and resources

The tables included in the Language WhitePapers provide an overview of the current situation of language technology support for different languages. The rating of existing tools and resources is shared by all Language WhitePapers and is based on educated estimations by several leading experts using the following criteria (each ranging from 0 to 6).

- Quantity: Does a tool/resource exist for the language at hand? The more tools/resources exist, the higher the rating.
  - 0: no tools/resources whatsoever
  - 6: many tools/resources, large variety
- Availability: Are tools/resources accessible, i.e., are they Open Source, freely usable on any platform or only available for a high price or under very restricted conditions?
  - 0: practically all tools/resources are only available for a high price
  - 6: a large amount of tools/resources is freely, openly available under sensible Open Source or Creative Commons licenses that allow re-use and re-purposing
- Quality: How well are the respective performance criteria of tools and quality indicators of resources met by the best available tools, applications or resources? Are these tools/resources current and also actively maintained?
  - 0: toy resource/tool
  - 6: high-quality tool, human-quality annotations in a resource
- Coverage: To which degree do the best tools meet the respective coverage criteria (styles, genres, text sorts, linguistic phenomena, types of input/output, number languages supported by an MT system etc.)? To which degree are resources representative of the targeted language or sublanguages?
  - 0: special-purpose resource or tool, specific case, very small coverage, only to be used for very specific, non-general use cases
  - 6: very broad coverage resource, very robust tool, widely applicable, many languages supported
- 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? Is the performance of such a technology adequate and ready for production use or is it only a prototype that cannot be used for production systems? An indicator may be whether resources/tools are accepted by the community and successfully used in LT systems.
  - 0: preliminary prototype, toy system, proof-of-concept, example resource exercise
  - 6: immediately integratable/applicable component
- Sustainability: How well can the tool/resource be maintained/integrated into current IT systems? Does the tool/resource fulfil a certain level of sustainability concerning documentation/manuals, explanation of use cases, front-ends, GUIs etc.? Does it use/employ standard/best-practice programming environments (such as Java EE)? Do industry/research standards/quasi-standards exist and if so, is the tool/resource compliant (data formats etc.)?
  - 0: completely proprietary, ad hoc data formats and APIs
  - 6: full standard-compliance, fully documented

- Adaptability: How well can the best tools or resources be adapted/extended to new tasks/domains/genres/text types/use cases etc.?
    - 0: practically impossible to adapt a tool/resource to another task, impossible even with large amounts of resources or person months at hand
    - 6: very high level of adaptability; adaptation also very easy and efficiently possible

2.1 Table of tools and resources for Bulgarian

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	4	3	5	5	4	3	4
Parsing (shallow or deep syntactic analysis)	2	2	4	4	3	3	3
Sentence Semantics (WSD, argument structure, semantic roles)	2	2	3	2	2	3	3
Text Semantics (coreference resolution, context, pragmatics, inference)	1	1	2	2	1	1	2
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	0	0	0	0	0	0	0
Information Retrieval (text indexing, multimedia IR, crosslingual IR)	2	1	2	2	2	2	3
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	2	1	3	3	2	2	3
Language Generation (sentence generation, report generation, text generation)	1	1	2	2	2	1	1
Summarization, Question Answering, advanced Information Access Technologies	2	2	2	2	2	1	2
Machine Translation	3	2	2	2	2	2	3
Speech Recognition	2	1	3	3	2	2	1
Speech Synthesis	2	1	3	3	2	2	1
Dialogue Management (dialogue capabilities and user modelling)	0	0	0	0	0	0	0
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	5	5	5	4	5	4	5
Syntax-Corpora (treebanks, dependency banks)	2	1	3	2	3	2	2
Semantics-Corpora	2	4	5	4	3	3	3
Discourse-Corpora	1	2	2	2	1	1	1
Parallel Corpora, Translation Memories	3	1	4	2	2	2	3
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	1	1	3	2	3	3	3
Multimedia and multimodal data (text data combined with audio/video)	1	1	1	1	1	1	1
Language Models	2	1	2	2	2	1	1
Lexicons, Terminologies	4	3	4	3	4	4	3
Grammars	2	2	3	3	3	3	2
Thesauri, WordNets	2	4	5	4	4	4	5
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	1	2	3	3	3	1	1

2.2 Table of tools and resources for Croatian

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	3	2	5	5	3	2	5
Parsing (shallow or deep syntactic analysis)	1	1	2	2	1	0	4
Sentence Semantics (WSD, argument structure, semantic roles)	1	0	1	3	0	0	3
Text Semantics (coreference resolution, context, pragmatics, inference)	0	0	0	0	0	0	0
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	0	0	0	0	0	0	0
Information Retrieval (text indexing, multimedia IR, crosslingual IR)	1	0	5	2	3	2	3
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	3	1	4	3	2	1	3
Language Generation (sentence generation, report generation, text generation)	1	1	4	0	3	0	0
Summarization, Question Answering, advanced Information Access Technologies	1	0	1	0	0	0	0
Machine Translation	1	0	1	3	0	0	0
Speech Recognition	2	2	3	3	2	3	3
Speech Synthesis	3	3	3	4	4	4	4
Dialogue Management (dialogue capabilities and user modelling)	1	2	1	1	0	2	2
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	3	3	3	4	4	4	2
Syntax-Corpora (treebanks, dependency banks)	1	1	3	4	2	1	2
Semantics-Corpora	0	0	0	0	0	0	0
Discourse-Corpora	0	0	0	0	0	0	0
Parallel Corpora, Translation Memories	3	2	3	3	3	1	2
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	3	1	3	3	4	3	4
Multimedia and multimodal data (text data combined with audio/video)	1	1	4	3	3	3	3
Language Models	0	0	0	0	0	0	0
Lexicons, Terminologies	3	3	4	3	4	3	3
Grammars	0	0	0	0	0	0	0
Thesauri, WordNets	2	3	3	4	3	2	2
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	0	0	0	0	0	0	0

2.3 Table of tools and resources for Hungarian

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	6	2	4	5	5	1	5
Parsing (shallow or deep syntactic analysis)	3	2	4	4	3	5	4
Sentence Semantics (WSD, argument structure, semantic roles)	1	3	3	1	0	0	3
Text Semantics (coreference resolution, context, pragmatics, inference)	1	3	2	0	0	0	3
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	0	0	0	0	0	0	0
Information Retrieval (text indexing, multimedia IR, crosslingual IR)	1	0	2	1	0	1	1
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	6	6	6	6	5	5	5
Language Generation (sentence generation, report generation, text generation)	0	0	0	0	0	0	0
Summarization, Question Answering, advanced Information Access Technologies	0	0	0	0	0	0	0
Machine Translation	6	1	5	5	6	5	6
Speech Recognition	3	0	4	2	4	3	3
Speech Synthesis	4	3	4	4	5	3	3
Dialogue Management (dialogue capabilities and user modelling)	0	0	0	0	0	0	0
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	6	6	6	6	6	6	4
Syntax-Corpora (treebanks, dependency banks)	1	6	6	5	6	6	4
Semantics-Corpora	3	6	6	1	3	5	5
Discourse-Corpora	0	0	0	0	0	0	0
Parallel Corpora, Translation Memories	6	4	6	6	6	6	6
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	2	2	4	2	4	4	0
Multimedia and multimodal data (text data combined with audio/video)	1	0	1	1	1	0	0
Language Models	6	3	4	3	6	6	5
Lexicons, Terminologies	5	1	6	6	2	2	6
Grammars	3	3	6	5	6	4	3
Thesauri, WordNets	1	1	6	3	5	5	3
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	2	6	1	1	1	4	2

2.4 Table of tools and resources for Polish

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	4	5	5	5	5	4	4
Parsing (shallow or deep syntactic analysis)	4	4	4	4	3	4	2
Sentence Semantics (WSD, argument structure, semantic roles)	1	2	4	1	1	2	4
Text Semantics (coreference resolution, context, pragmatics, inference)	1	0	3	1	0	1	1
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	1	0	1	1	0	1	0
Information Retrieval (text indexing, multimedia IR, crosslingual IR)	2	5	2	2	4	4	4
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	3	3	2	2	1	3	4
Language Generation (sentence generation, report generation, text generation)	1	1	1	1	1	1	2
Summarization, Question Answering, advanced Information Access Technologies	1	1	2	2	1	1	1
Machine Translation	3	4	3	3	3	4	3
Speech Recognition	1	2	3	4	3	2	4
Speech Synthesis	4	3	6	5	4	4	3
Dialogue Management (dialogue capabilities and user modelling)	1	1	1	1	1	1	1
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	3	2	4	4	5	5	3
Syntax-Corpora (treebanks, dependency banks)	2	1	4	4	1	4	4
Semantics-Corpora	1	0	4	2	2	0	4
Discourse-Corpora	1	1	2	1	1	1	1
Parallel Corpora, Translation Memories	3	1	4	4	5	5	5
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	1	0	3	3	2	2	2
Multimedia and multimodal data (text data combined with audio/video)	1	1	1	1	1	0	0
Language Models	1	3	1	1	1	1	1
Lexicons, Terminologies	3	2	4	4	4	4	2
Grammars	3	2	4	4	3	2	2
Thesauri, WordNets	3	4	4	4	3	4	4
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	2	1	4	2	1	1	2

2.5 Table of tools and resources for Serbian

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	4	3	5	5	5	4	4
Parsing (shallow or deep syntactic analysis)	1	2	5	3	2	2	2
Sentence Semantics (WSD, argument structure, semantic roles)	0	0	0	0	0	0	0
Text Semantics (coreference resolution, context, pragmatics, inference)	0	0	0	0	0	0	0
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	0	0	0	0	0	0	0
Information Retrieval (text indexing, multimedia IR, crosslingual IR)	3	1	3	3	2	2	3
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	1	2	2	2	3	2	3
Language Generation (sentence generation, report generation, text generation)	0	0	0	0	0	0	0
Summarization, Question Answering, advanced Information Access Technologies	1	1	0	1	0	1	1
Machine Translation	1	1	0	1	0	1	1
Speech Recognition	2	2	1	1	1	1	0
Speech Synthesis	2	2	4	4	5	5	1
Dialogue Management (dialogue capabilities and user modelling)	0	0	0	0	0	0	0
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	2	4	2	4	4	4	4
Syntax-Corpora (treebanks, dependency banks)	0	0	0	0	0	0	0
Semantics-Corpora	0	0	0	0	0	0	0
Discourse-Corpora	0	0	0	0	0	0	0
Parallel Corpora, Translation Memories	3	3	3	2	2	2	3
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	1	2	4	4	3	3	3
Multimedia and multimodal data (text data combined with audio/video)	1	2	2	1	2	1	2
Language Models	1	3	2	3	2	2	3
Lexicons, Terminologies	2	3	4	4	3	3	3
Grammars	1	1	0	1	0	1	1
Thesauri, WordNets	2	4	3	2	4	2	4
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	1	1	0	1	0	1	1

2.6 Table of tools and resources for Slovak

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	2	2	3	4	4	3	3
Parsing (shallow or deep syntactic analysis)	0	0	0	0	0	0	0
Sentence Semantics (WSD, argument structure, semantic roles)	0	0	0	0	0	0	0
Text Semantics (coreference resolution, context, pragmatics, inference)	0	0	0	0	0	0	0
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	0	0	0	0	0	0	0
Information Retrieval (text indexing, multimedia IR, crosslingual IR)	3	1	2	3	4	2	1
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	1	4	1	1	1	2	2
Language Generation (sentence generation, report generation, text generation)	0	0	0	0	0	0	0
Summarization, Question Answering, advanced Information Access Technologies	1	2	1	1	1	3	3
Machine Translation	2	2	2	2	2	1	2
Speech Recognition	3	1	2	2	3	3	2
Speech Synthesis	3	3	3	3	3	3	3
Dialogue Management (dialogue capabilities and user modelling)	0	0	0	0	0	0	0
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	2	4	4	5	4	4	4
Syntax-Corpora (treebanks, dependency banks)	1	4	2	2	2	3	3
Semantics-Corpora	0	0	0	0	0	0	0
Discourse-Corpora	1	1	2	1	3	2	3
Parallel Corpora, Translation Memories	2	3	2	2	2	2	3
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	3	4	2	2	3	3	3
Multimedia and multimodal data (text data combined with audio/video)	1	1	3	2	2	3	3
Language Models	1	4	1	3	3	3	4
Lexicons, Terminologies	3	2	3	4	3	4	1
Grammars	3	3	3	2	1	2	1
Thesauri, WordNets	2	5	2	1	2	4	4
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	0	0	0	0	0	0	0

3. General summaries for the languages:

Key results regarding technologies and resources include the following:

3.1. Bulgarian

For morphologically related tools such as tokenizers, part of speech taggers and morphological analyzers, the situation in Bulgaria is reasonably good. Even if the tools are not all freely available, the resources are of relatively high quality and the coverage is good.

With regard to resources such as reference corpora, lexicons, and wordnets, the situation is also reasonably good for Bulgarian since substantial resources have been developed in recent years. But while some reference corpora of high quality and quantity exist, i.e. the Bulgarian National Corpus, large syntactically and semantically corpora annotated by experts are not available.

Semantic tools and resources are scored very low. Thus, programs and initiatives are needed to substantially boost this area both with regard to basic research and the development of annotated corpora.

There also exist individual products with limited functionality in subfields such as speech synthesis, speech recognition and machine translation, and a few others.

There are insufficient parallel corpora for machine translation. Translation of Bulgarian into and from another language works best since most data exists.

There is a huge gap in multimedia data.

Several of the resources lack standardization, i.e., even if they exist, sustainability is not supported; concerted programs and initiatives are needed to standardize data and interchange formats.

3.2. Croatian

Croatian stands reasonably well with respect to the most basic language technology tools and resources, such as reference corpora, smaller parallel corpora, large inflectional lexicons, tokenisers, MSD taggers, lemmatisers, NERC system etc.

However, a large syntactically annotated corpus is missing as well as large parallel corpus (e.g. Croatian translations of Acquis Communautaire).

Many of existing resources lack standardization so initiatives are needed to standardize the data and interchange formats.

Experiments have been conducted in some areas, such as shallow parsing (chunking), summarization, application of ontological resources, but only in an academic research environment. However, the results obtained are far from the level of development that other European languages demonstrate.

The multimedia and multimodal document processing, is gaining attraction, particularly the digitisation in the context of preserving the cultural heritage.

There exist also individual products with limited functionality in subfields such as speech synthesis, speech recognition and information extraction, and a few others.

Tools and resources for more advanced language technology such as deep parsing, machine translation, text semantics, discourse processing, language generation, dialogue management, etc., simply do not exist.

3.3. Hungarian

While some specific corpora of high quality exist, a very large syntactically annotated corpus is not available.

There is one syntactically highly elaborately annotated corpus for Hungarian. The corpus is available for free, which results in a wide range applications developed based on the corpus.

Many of the resources lack standardization, i.e., even if they exist, sustainability is not addressed; concerted programs and initiatives are needed to standardize data and interchange formats.

Standards do exist for semantics in the sense of world knowledge (RDF, OWL, etc.); they are, however, not easily applicable to NLP tasks.

The standard preprocessing steps (tokenization, morphology, shallow parsing etc.) are complete for Hungarian, but the more difficult semantics and discourse analysis need further research.

Research was successful in designing particular high quality software, but it is nearly impossible to come up with sustainable and standardized solutions given the current funding situations.

There is a large variation in the different areas of NLP in Hungarian: in some fields (e.g. morphology, IE, MT, parallel corpora) there are higher ratings, while in other fields (e.g. advanced discourse processing, dialogue management) ratings are close to zero.

Speech Recognition and Machine Translation of Hungarian is studied at several universities and workplaces, but free tools and data are not available. It is quite typical in the Hungarian NLP market: the number of free databases and open source programs is quite low.

3.4. Polish

For Polish, discourse corpora or advanced discourse processing are not widely available. Multimodal corpora are in preparation.

Many of the resources lack standardization, i.e. even if they exist, sustainability is not given; concerted programs and initiatives are needed to standardize data and interchange formats.

Semantics is more difficult to process than syntax; text semantics is more difficult to process than word and sentence semantics.

Standards do exist for semantics in the sense of world knowledge (RDF, OWL, etc.); they are, however, not easily applicable to NLP tasks.

Speech processing, specially speech synthesis, is currently more mature than NLP for written text.

Research was successful in designing particular high quality software, but it is nearly impossible to come up with sustainable and standardized solutions given the current funding situations.

Polish lacks large, balanced and more easily available parallel corpora, including large parallel corpora for related languages such as Czech or Polish.

For many purposes, bilingual and multilingual dictionaries that include not only translations but also valency information seem indispensable. These need to be built, as standard dictionaries usually omit this kind of annotation.

Large and widely available ontological resources for Polish are needed for many applications. Currently available ontologies are relatively small, based on OpenCyc or on Polish OpenThesaurus. A Polish version of DBPedia is in preparation.

3.5. Serbian

When morphological issues and issues related to them are concerned, it is safe to say that the level of development of technologies and resources is satisfactory, mainly due to the existence of large electronic dictionaries and local grammars. An immediate consequence of this fact is that necessary tools for information retrieval and information extraction are available. Some of the dictionaries are ready for wider use, whereas some need to be upgraded, as for example SrbNet.

A reference corpus of contemporary Serbian in ekavian dialect is available, as well as several parallel aligned corpora, all of which are available to researchers of Serbian. Current research is focused on upgrading of the reference corpus and its expanding with the ijekavian dialect.

Speech technologies are well developed, and they have found wide usage in business, but research needs to be further expanded, in order to expand the area of their usability.

Software aimed at enhancing the productivity of lexicographical work has been developed, but the issue of accepting new technologies in traditionally oriented lexicographic environments is an impediment to speedier development of lexicography.

Successful experiments have been performed in some areas, such as shallow parsing, summarization, machine translation, ontological resources, in a strictly research environment. However, the results obtained are still far from the level of development reached for developed European languages. The attention of researchers is also attracted by the multimedia and multimodal document, especially in the context of digitization of cultural heritage.

Given the complexity of Serbian syntax, areas based on deep parsing simply do not exist: sentence semantics, text semantics, and language generation. This results in the absence of a formalized syntax of Serbian and restricts the development of syntactically and semantically annotated corpora. Formalization of Serbian syntax is thus the most urgent task for further expansion of HLT.

3.6. Slovak

For Slovak, the Slovak National Corpus is the reference language corpus, but only the query interface is generally available, due to licensing restrictions.

On the other hand, Corpus of Spoken Slovak is not encumbered by copyright law and is therefore publicly available, but its size is miniscule compared with the corpus of written language.

While some specific corpora of high quality exist, a very large syntactically annotated corpus is not available.

Many of the resources lack standardization, i.e., even if they exist, sustainability is not given; concerted programs and initiatives are needed to standardize data and interchange formats.

There is an ontological resource for Slovak (even mapped to English ontological resources) but its coverage is limited.

Standards do exist for semantics in the sense of world knowledge (RDF, OWL, etc.); they are, however, not easily applicable to NLP tasks.

Many of the resources taken as standard in other languages are missing for Slovak; NLP language research in Slovakia is severely underfunded.

Some of the research and development activities for the Slovak language is carried out in the Czech Republic by Czech universities and Czech SMEs.

Speech Recognition of the Slovak language is studied at several universities and workplaces but the amount of free tools and data is limited. In contrast with speech recognition, speech synthesis is less covered by universities and other workplaces. In the field of speech synthesis, there are open source packages available together with several other simple synthesizers but the speech synthesis with more natural voices is not available.

Slovak dialogue systems are very little extended due to poor accessibility of high quality speech recognition modules of the Slovak language.

4. Brief general conclusions

Public funding for LT in Europe is relatively low compared to the expenditures for language translation and multilingual information access by the USA. In general in Central and South-East European countries public funding is even lower than in many other European countries. Although there is a pressing need of recognising the importance of LT in ensuring sustainable development of our languages in 21st century, a limited number of initiatives has been launched, that would foster the creation of large-scale resources and tools/services, as well as a partnership between government, academia and industry to develop an expertise cluster in natural language technology. We believe that such initiatives should be institutionally supported in order to stimulate business research and promote sectoral cooperation between companies and research institutions to develop innovative products and technologies to improve the competitiveness of enterprises at EU market.

5. Annexes

Bulgarian Language WhitePaper
Bulgarian Language WhitePaper in Bulgarian
Croatian Language WhitePaper
Croatian Language WhitePaper in Croatian
Hungarian Language WhitePaper
Hungarian Language WhitePaper in Hungarian
Polish Language WhitePaper
Polish Language WhitePaper in Polish
Serbian Language WhitePaper
Serbian Language WhitePaper in Serbian
Slovak Language WhitePaper
Serbian Language WhitePaper in Slovak