CESAR

Central and South-East European Resources

Project no. 271022

Deliverable D2.4
Report on methodology and criteria for the selection of resources

Version No. 1.2

03/11/**2011**

Document Information

Deliverable number:	2.4
Deliverable title:	Report on methodology and criteria for the selection of resources
Due date of deliverable:	30 September 2011
Actual submission date of deliverable:	03 Novmber 2011
Main Author(s):	Svetla Koeva (IBL)
Participants:	Maciej Ogrodniczuk, Adam Przepiórkowski (IPIPAN) Piotr Pęzik (ULodz) Eszter Simon (RILHAS), György Szaszák (BME) Rada Vlahova (IBL) Radovan Garabík (LSIL) Duško Vitas, Cvetana Krstev (UBG) Marko Tadić, Ida Raffaelli (FFZG)
Internal reviewer:	Tamás Váradi (RILHAS)
Workpackage:	2
Workpackage title:	Analysis and selection of language resources
Workpackage leader:	IBL
Dissemination Level:	Public
Version:	1.2
Keywords:	language resources, methodology for the selection of resources

History of Versions

Version	**Date**	**Status**	**Name of the Author (Partner)**	**Contributions**	**Description/ Approval Level**
1.2	03/11/2011	Modification	Svetla Koeva (IBL)	modifications	Modification upon the request of IPIPAN
1.1	30/09/2011	Final version	Eszter Simon Tamás Váradi (RILHAS)	proofreading	Revision
1.0	29/09/2011	Draft	Svetla Koeva (IBL)	draft	Draft of first version based on contributions from partners listed above (Participants)

EXECUTIVE SUMMARY

The Deliverable 2.4. *Report on methodology and criteria for the selection of resources* describes the methodology and criteria that allow partners to assess the quality and importance of language resources and tools, thus enabling thum to set appropriate priorities. The aim is to ensure a balanced coverage of selected resources and tools for different end-users and tasks, groups of products and services.

Table of Contents

1. Introduction

The Deliverable 2.4. *Report on methodology and criteria for the selection of resources* reports on the results from the accomplishment of the Task 2.3 *Selection of resources of further interest within* the CESAR WP2. All partners have participated in the task. The report describes the methodology and criteria that are used for a precise selection of resources and tools.

Methodology and criteria that allow partners to assess the quality and importance of language resources and tools are established, thus enabling the CESAR project partners to set appropriate priorities. The aim is to ensure a balanced coverage of resources and tools for different end-users and tasks, groups of products and services.

On the basis of the agreed methodology and criteria, the consortium selected the best possible mix of resources that will further raise the interest of different groups of end-users. The outcome of this task allows CESAR partners to provide an analysis of the current situation and make suggestions in case of lack of essential resources for covered languages to determine further efforts of the community.

Chapter 2 introduces the adopted methodology and criteria for language resource selection. Criteria of quality assessment are proposed with particular attention to current developments. The approach for language resources and tools selection is based on a list of indicators, where each language resource is specified according to different groups of criteria. The goal is to ensure as accurate measurement as possible for different quality and quantity parameters.

Chapter 3 illustrates how the adopted methodology and criteria are applied for each individual language. It gives an overview of the existing language resources for the every language and makes it possible to identify the gaps in the provision of the language resources and tools components. For each language, an analysis of the set of already selected resources and tools is provided, together with an outline of the gaps. The analysis leads to the conclusion what kind of resources should be of the greatest interest in next rounds of selection.

The report ends with general conclusions for the application of methodology and criteria for selection of resources.

2. The methodology and criteria adopted for resource selection

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 as follows:

- **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 a certain 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 thorough analysis.

- **Total Point Value**

Following the approach of the EU NEMLAR (Network for Euro-Mediterranean Language Resources) project (concerning a BLARK for Arabic), 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

i.1. Available for whom?

company-internal (3)
freely usable for PreR&D (2)
freely usable for both PreR&D and R&D (1)

i.2. At what price?

Over 10K EUR (4)
Between 1 and 10K EUR (3)
Between 100 EUR and 1K EUR (2)
Less than 100 EUR or free (1)

i.3. How straightforward it is to reuse it (degree of adaptability)?

Black box resource (3)
Glass box resource (2)
Open resource (1)

ii. Quality

ii.1. Standard compliance (Is the resource based on a common standard?)

No common standards used (3)
No common standards used internally, but interfaces or converters to standards are available (2)
Standard-compliant (1)

ii.2. Soundness (Internal consistency, i.e., is the resource based on well-defined specifications?)

No specifications available (3)
Specifications cover only certain aspects of tools (2)
Full specification exists (1)

ii.3. Task-relevance (Is the resource suited for a specific task?)

Not particularly well-suited, should be improved (3)

To a certain extent (2)
Very well suited (1)

ii.4. Environment-relevance (Is the resource interoperable with other resources?)
No (3)
Yes, with a limited number of them (2)
Yes, with many of them (1)

iii. Quantity (resources only)
Below 50 per cent of top quantity available for the language (3)
Between 50 and 90 per cent of top quantity (2)
Over 90 per cent of top quantity (1)

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.

- **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.
0: no tools/resources whatsoever
6: many tools/resources, large variety

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?
0: practically all tools/resources are only available at a high price
6: a large amount of tools/resources is freely and openly available

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?
0: toy resource/tool
6: high-quality tool, human-quality annotations in a resource

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?
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

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?
0: preliminary prototype, toy system, proof-of-concept, example resource exercise
6: immediately integratable / applicable component

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

0: completely proprietary, ad hoc data formats and APIs

6: full standard-compliance, fully documented

vii. **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

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.

- **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.

- **BLARK - the minimum set of resources**

The BLARK (Basic Language Resources Kit) concept was defined by a joint initiative between ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association). BLARK is defined *as the minimal set of resources that is necessary to do any precompetitive research and education at all*. The 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 (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.

The 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 agains and contra selection of a particular resource. That is way, despite our initial intention, the 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 criteria **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 quantitate 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.

3. Application of the adopted methodology for each of the individual languages

3.1. Bulgarian

3.1.1. General evaluation of Bulgarian resources

For upgraded resources:

- **All selected resources are state-of-the-art representatives of their type for Bulgarian:**
  - Bulgarian National Corpus: a publicly available constantly enlarged corpus (app. half a billion words) designed as a uniform environment for texts of different modality, period, and language
  - Bulgarian-X+ Language parallel corpus: its largest part, Bulgarian-English parallel corpus consists of app. 100M words per language
  - Bulgarian PoS-annotated corpus: the largest manually morphologically annotated corpus for Bulgarian
  - Bulgarian Sense-annotated corpus: the only manually sense annotated (with wordnet senses) corpus for Bulgarian
  - Data base of spoken Bulgarian: the largest freely available Bulgarian spoken corpus
  - Bulgarian morphological dictionary: large inflexional dictionary of Bulgarian based on the latest official orthography dictionary
  - Bulgarian frequency dictionary: based on the very large Bulgarian National Corpus
  - Bulgarian wordnet: the only wordnet for Bulgarian and one of the biggest in Europe
  - Bulgarian Framenet: the only Framenet for Bulgarian and the largest frame dictionary for Bulgarian
- **Equally valuable representatives are included in the selection**
  - Data base of spoken Bulgarian includes practically many of the available spoken resources
  - Parallel corpora represent different domains following the structure of the Bulgarian National Corpus
- **Some of the representatives are the only available for Bulgarian on that type**
  - Bulgarian Sense-annotated corpus: the only manually Sense-annotated corpus for Bulgarian
  - Bulgarian wordnet: the only wordnet for Bulgarian
  - Bulgarian Framenet: the only Framenet for Bulgarian
- **Current status of resources present superior quality at least on regional level without the need of excessive further development:**
  - All resources are among the best available resources for Bulgarian
  - The following resources are ready to be upgraded without any development: Bulgarian morphology dictionary, Bulgarian PoS and Sense-annotated corpora; Bulgarian frequency dictionary
- **Licensing issues allow to freely process and make available the resources and resource-related materials or the consortium succeeds in reaching an agreement with respective copyright holders:**
  - Most of the resources are either publicly available PUB / CC BY or via the META-SHARE ACA NC. Some of the resources have been already distributed

by ELDA. At the moment Bulgarian National Corpus, Bulgarian Sense-annotated Corpus and Bulgarian morphological dictionary are available for online search for research and education purposes.

For extended/linked resources:

- **The extension of resources provides considerable value to the community, at least on regional level:**
  - Bulgarian X-Language parallel corpus: increasing the source data significantly improves its range of usage
  - Bulgarian wordnet: extension the number of synsets improves its range of applications
  - Bulgarian National Corpus: adding higher level annotation makes them applicable for wide range of linguistic analysis and natural language processing tasks; providing a web service for collocations' statistics will allow the corpus to be used for different NLP tasks
  - Bulgarian-X language parallel corpus: providing online search interface and web service for translation equivalents search will allow the corpus to be used for different NLP tasks
  - Bulgarian morphological dictionary: its availability as Bulgarian Spell Checker for Windows, Mac OS and web service allows the dictionary to be used widely
  - Bulgarian processing components (splitter, tokenizer, tagger and lemmatizer): linking Bulgarian processing components into a tool chain improves their adaptability and maturity

- **The emphasis is on providing building blocks to the existing tools rather than major restructuring**:
  - All of the planned actions are aiming in additions of new resource units rather than in restructuring of the resources: words, synsets, frames, source data, higher level of annotation, etc.

- **Additional resources are integrated with existing ones only if they significantly improve the quality of resulting resources**:
  - Existing dictionaries of proper names, abbreviations, multiword expressions will be integrated with the Bulgarian morphology dictionary

- **If more than one representative of certain tool type for a language has been selected, they are very likely to be interlinked to benefit from strong points of both solutions**:
  - Data from Bulgarian wordnet, Bulgarian Framenet and Bulgarian morphological dictionary will be integrated in order to keep resources consistence and to support common conventions
- I**f less-developed, but still very popular tools can benefit from the enhancement basing on their well-developed equivalent, their enhancement is also considered**:
  - Frequency dictionary can be constantly adjust according to the data from the permanently enlarging corpora

- **Experience of other consortium members/other consortia is extensively used in the process of further extending national resources to provide strong foundation for cross-linguality**:

- Selection of resources and tools are based on consultation with partners both inside and outside the consortium

- **Tools offering language-neutrality or cross-linguality are preferred**:
  - Some of the tools contain language-independent modules
  - Several resources follow international standards in annotation

For resources aligned across languages:

- **No more than one tool of a certain type for each language is used:**
  - Bulgarian X-language parallel corpus: providing cross-language alignment for different languages
  - Bulgarian wordnet: aligning synsets to PWN and thus across similar resources within the project which are aligned in a similar way
  - Bulgarian Framenet: aligning frames to Framenet and thus across similar resources within which are aligned in a similar way
- **Bulgarian Language Processing Tools in NooJ**:
  - Comparable with the modules for other project's languages
- **Whenever applicable, the largest set of languages is selected**:
  - Most of the available parallel resources (OPUS, EuraLex, SETimes, JRC Acquis, etc.) are gathered
- **Language-independence is targeted to a great extent**:
  - Language-independence is obtained on the representation format level
  - Some of the modules of the language processing tolls are language independent
- **The quality of a result is of immense concern**:
  - The quality of the results is maintained by compliance with agreed standards on all levels of annotation

3.1.2. Total Point Value for Bulgarian resources

The TPVs were calculated for the resources and tools being evaluated according to measures specified in CESAR Description of Work.

Resource	Available for whom?	For how much?	Adapta-bility	Comp-liance	Soundness	Task-relevance	Environment-relevance	Quantity	TPV
Bulgarian National corpus	1	1	1	1	1	1	1	1	**8**
Bulgarin X-language parallel corpus	2	1	1	1	1	1	1	2	**10**
Bulgarian PoS-annotated corpus	2	1	1	1	1	1	1	2	**10**
BG Sense-annotated corpus	2	1	1	1	1	1	1	1	**9**
Bulgarian wordnet	1	2	1	1	1	1	1	1	**9**
Bulgarian FrameNet	3	2	1	1	1	1	1	1	**11**
Bulgarian morphology dictionary	2	2	1	1	1	1	1	1	**10**

Data base of spoken Bulgarian	1	1	1	1	1	1	1	2	**9**

Table 1: The TPVs for Bulgarian resources

As the table shows, resources and tools with a certain level of TPV (bellow 16) were selected. As mentioned above the selection was based also on the criteria described in the first section, thus the importance, usability and popularity within the community were important factors.

3.1.3. How Bulgarian LWP reflects on the selection

The Bulgarian edition of the Language White Paper is also a useful source for the evaluation and comparing available language resources in various groups of potential applicability:

	Quantity	Availability	Quality	Coverage	Maturity	Sustaina-bility	Adaptability
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 SD, annotated SD, dialogue SD)	1	1	3	2	3	3	3
Multimedia and multimodal data (text with AV)	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	1	2	3	3	3	1	1

Table 2. Data from the Bulgarian White Paper

The results of the table are useful for showing where real gaps in Bulgarian language resources can be detected. At first place multimedia and multimodal data resources (texts

combined with audio and / or video) are missing - it will be hard to fill this gab in the scope of the project since the reports for data collections are very rare). Speech corpora are the next important target - the available speech corpora are of the good quality but much more data are needed for any ambitious research or data processing system.

3.1.4. Proportion between the selected resources for Bulgarian developed inside and outside the consortium

14% of the selected resources and tools for Bulgarian were developed outside the consortium, 86% of them were developed inside the consortium:

Resource name	Developed inside/outside the consortium?
Bulgarian National corpus	internal
Bulgarin X-language parallel corpus	internal
Bulgarian PoS annotated corpus	internal
Bulgarian Sense annotated corpus	internal
Bulgarian wordnet	internal
Bulgarian FrameNet	internal
Bulgarian morphology dictionary	internal
Data base of spoken Bulgarian	external

Table 3. Bulgarian resources developed inside and outside the consortium

3.1.5. Analysis of the set of the already selected Bulgarian resources and tools

Regarding its technological impact, CESAR targets specific Bulgarian language resources with a view to improving their availability, interoperability and representativeness. The following general categories of resources strictly follow this requirement:

- Morphological dictionary and annotated corpora – a basic prerequisite for most NLP solutions.
- Parallel corpora annotated and aligned in widely accepted standards.
- Higher-level syntactic and semantic resources such as Wordnets, Framenets for Bulgarian.
- Language processing tools from tokenization to higher level processing.

Dictionaries

Dictionaries are basic components in Natural Language Processing. 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 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 are as follows: the Bulgarian morphological dictionary 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 Bulgarian spell checker - one of the targets for a further distribution within the CESAR.

Written corpora

The Bulgarian National Corpus is compiled mainly for the purposes of computational research and implementations, and the same is the function of the parallel corpora within. At the present moment the Bulgarian National Corpus has app. half a billion words. Every document is accompanied with extended meta-textual information in XML format. The unified description of texts facilitates their processing and grouping in relevant subcorpora on the basis of various criteria. The corpus is automatically annotated for part of speech, grammatical characteristics, lemma and wordnet word senses. The Bulgarian National Corpus is a language resource of national importance and provides a wide range of possibilities for theoretical and practical applications in a number of areas. Since mid 2009 the Bulgarian National Corpus has been publicly accessible on the internet.

Parallel corpora

Parallel corpora are among the most important resources used in multilingual language processing. They serve as training data sets in machine translation systems, cross-linguistic information retrieval and in the construction of bilingual dictionaries. A brief overview of parallel corpora developed so far, where Bulgarian is one of the languages in focus, gives reasons to conclude that those corpora are not very extensive; they represent generally administrative or literally texts and they are built from the available texts on the internet, rather than on a planned strategy for developing a balanced and representative parallel corpus. On the other hand, the selected Bulgarian X-language is constantly enlarging both in amount of documents and in number of languages; it is organised following the structure of the Bulgarian National Corpus, and provided with the same kind of annotation and metadata description. Thus any X-language from the parallel texts is equally treated with respect to the text type diversity and balance, metadata description scheme, preprocessing and annotation, search engine queries and database, user interface and visualisation of results.

Annotated corpora

Manually annotated corpora are important resources, used for training and testing various language processing tools. They also provide a training data and a reference to measure the performance of tools with the same function. Two manually annotated corpora are selected so far - Bulgarian PoS-annotated and Sense-annotated corpora. In the Bulgarian POS-annotated Corpus (+150000 words) each word form is annotated by hand with the relevant part of speech and grammatical features, with which it is used in the context, selected from a majority of possibilities from the large Bulgarian morphological dictionary. In the sense-annotated corpus (+100000 words) each lexical unit is linked manually with the most appropriate synonym set from the Bulgarian wordnet. Unlike the bulk of sense-annotated corpora where only (sets of) content words are annotated, in Bulgarian Sense-annotated Corpus each lexical unit has been assigned a sense.

Speech databases

Representative and phonetically balanced speech corpora are among the most important resources. Data base of Spoken Bulgarian consists of Interview, Media and Formal speech, Student's speech, Academic speech and Colloquial speech.

Wordnets and Framenets

Wordnet and FrameNet undoubtedly occupy an important place amongst lexical resources which have been very important for the creation of more complex applications in the area of Natural Language Processing. The Bulgarian wordnet is one of the most complete and consistent lexical resources (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 different languages are connected by means of inter-language equivalence relations, which are used as a basis for the development of the wordnet multilingual lexical-semantic network, the so collet global wordnet. The Bulgarian wordnet is approximately one quarter the size of the English wordnet and is one of the biggest in Europe.

The Bulgarian FrameNet represents general semantic and language-specific lexical-semantic and syntactic combinatory properties of Bulgarian lexical units. The unique character of the Bulgarian FrameNet is determined by the fact that it defines classes of lexical units in relation to: their place in a given semantic frame at an inter-language level, their productivity in the formation of diathesis, semantic and syntactic alternations, the expression of general morpho-syntactic characteristics and the description of (combinations of) obligatory and permissible environments.

With regard to resources such as lexicons, Wordnets and Framenets in Bulgaria substantial resources have been developed in recent years, although their enlargement and cross-validation are subject of further work.

Language processing tools

The standard preprocessing steps (tokenization, morphological analysis, POS tagging, lemmatizer) are available for Bulgarian. The Bulgarian language processing chain is designed for integration in different NLP applications: it does not change the input texts and can be easily adjusted to various inputs; the annotation is accumulated rather than substituted or transformed; all modules have a C++ implementation and thus are platform independent. Some of the resources are used for training and testing corpora in the process of developing the tools (Bulgarian PoS-annotated and Sense-annotated corpora).

3.1.6. Bulgarian resources of greatest interest in the next rounds of selection

Parallel corpora have been given one of the highest priorities in the upgrade, extension and alignment process, so more resources of this type are intended to be included in the next batches of resources (to be released in June 2012 and January 2013) or even going beyond the end of the project. Multimedia and multimodal resources are not available - it will be hard to fill this gab in the scope of the project. Speech corpora are the next important target - the available speech corpora are of the good quality but much more data are needed for any ambitious project for processing Bulgarian speech.

3.2. Croatian

3.2.1. General evaluation of Croatian resources

For upgraded resources:

- **All selected resources are state-of-the-art representatives of their type for Croatian:**
- Croatian National Corpus: the first Croatian 100-million corpus covering standard Croatian written, different genres, domains, text types
- Croatian-English Parallel Corpus: the largest Croatian-English parallel corpus
- Croatian wordnet: the only wordnet for Croatian
- Southeast European Parallel Corpus: parallel and strongly comparable corpus in 10 languages from Southeast Europe
- Croatian Dependency Treebank: the only treebank for Croatian
  - Croatian Morphological Lexicon: the largest Croatian inflectional lexicon
  - Croatian Valency Dictionary (CROVALLEX): the only Croatian valency dictionary

  - Croatian Language Corpus: the largest diachronic Croatian text collection
- **Equally valuable representatives are all included in the selection:**
  - Croatian webcorpus: the largest Croatian corpus
  - Croatian Language Processing Tools in NooJ
  - Croatian NERC system
  - Croatian Lemmatization Server
  - CollEx: language independent collocation extractor
- **Current status of resources present superior quality at least on regional level without the need of excessive further development:**
  - All resources are the best available resources for Croatian; all are ready to be upgraded without much additional development (except Croatian wordnet that needs moderate upgrading to reach over 10,000 synsets)
- **Licensing issues allow to freely process and make available the resources and resource-related materials or the consortium succeeds in reaching an agreement with respective copyright holders:**
  - Most of the resources will be available using META-SHARE ACA NC. Also Croatian National Corpus, Croatian Morphological Lexicon/Croatian Lemmatisation Server are available for online search using client-based or web-based interface available for research and education.

For extended/linked resources:

- **The extension of resources provides considerable value to the community, at least on regional level:**
  - Croatian-English Parallel Corpus and Southeast European Parallel Corpus: increasing the source data significantly improves its range of usage
  - Croatian wordnet: extension the number of synsets improves its range of application
  - Croatian National Corpus and Croatian webcorpus: adding additional level annotation makes them applicable for wide range of linguistic analysis
  - Croatian NERC system and Croatian Language Processing Tools in NooJ: extension to new domains increases their coverage and adaptability
  - Croatian Lemmatization Server: linking the web service into a tool chain with a standardized API improves their adaptability and maturity
- **The emphasis is on providing building blocks to the existing tools rather than major restructuring**:
  - All of the planned actions are additions of synsets, source data, higher level annotation, new entries etc., along the existing structure and balance of different sources within existing resources and it does not include restructuring of existing resources
- **Additional resources are integrated with existing ones only if they significantly improve the quality of resulting resources**:
  - No additional resources are planned at this stage of the project. However, if some new, valuable resource appears outside of the project, we will try to adopt it within the CESAR and META-SHARE framework
- **If more than one representative of certain tool type for a language has been selected, they are very likely to be interlinked to benefit from strong points of both solutions**:
  - So far for Croatian there are no overlapping in existing resources and/or tools
- **If less-developed, but still very popular tools can benefit from the enhancement basing on their well-developed equivalent, their enhancement is also considered**:
  - No tools involved

- **Experience of other consortium members/other consortia is extensively used in the process of further extending national resources to provide strong foundation for cross-linguality**:
  - Selection of resources and tools are based on consultation with partners both inside and outside the consortium
- **Tools offering language-neutrality or cross-linguality are preferred**:
  - Some tools are designed as language-independent modules (CollEx), while Croatian language-dependent tools and resources follow international or de facto standards in annotation (e.g. TEI guidelines for corpora, MulTextEast lexica format for Croatian Morphological Lexicon, NooJ format of grammars and lexica)

For resources aligned across languages**:**

- **No more than one tool of a certain type for each language is used**:
  - Croatian-English Parallel Corpus: providing cross-language alignment
  - Croatian Wordnet: aligning synsets across similar resources within the project (plWordNet, Serbian Wordnet, Bulgarian Wordnet, Hungarian Wordnet)
  - Croatian Language Processing Tools in NooJ: addition modules for project's languages
  - CollEx: extension of language model to work with other project languages
  - NERC: local grammars to recognise named entities in other project languages could be developed
- **Whenever applicable, the largest set of languages is selected**:
  - All potentially available parallel resources were gathered (particularly Southeast European Parallel Corpus)
- **Language-independence is targeted to a great extent**:
  - Language-independence is obtained on the representation format level; the data for parallel corpora is language-dependent by default
- **The quality of a result is of immense concern**:
  - The quality of the result is maintained by compliance with agreed standards on all levels of annotation; special effort has been made to make the parallel corpora available not only with the research community formats (TEI), but also in an industry-wide format (TMX)

3.2.2. Total Point Value for Croatian resources

The selection process was not based on TPV factors, but the criteria described above. Nevertheless, TPVs were calculated for the resources and tools being evaluated according to measures specified in CESAR:

Resource	Available for whom?	For how much?	Adap-tability	Comp-liance	Soundness	Task-relevance	Environment-relevance	Quantity	TPV
Croatian National corpus	1	1	2	1	1	1	2	1	10
Croatian-English PC	1	2	2	1	1	1	1	1	10
Southeast European PC	1	1	1	2	3	3	2	2	15
Croatian Dependency Treebank	2	3	2	1	2	1	1	1	13

Croatian Morphological Lexicon	**1**	**3**	**2**	**1**	**1**	**1**	**1**	**1**	**11**
Croatian wordnet	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**1**	**15**
Croatian Valency Dictionary	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**9**
Croatian Language Corpus	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**2**	**13**
Croatian Web Corpus	**2**	**1**	**1**	**2**	**2**	**1**	**2**	**1**	**12**
Croatian LPTs in NooJ	**3**	**1**	**2**	**2**	**1**	**2**	**2**	**0**	**13**
Croatian NERC system	**3**	**3**	**3**	**2**	**2**	**1**	**2**	**0**	**16**
CollEx	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**0**	**8**

Table 4. The TPVs for Croatian resources

As the table shows, resources and tools with a certain level of TPV (bellow or equal to 16) were selected. As mentioned above the selection was based also on the criteria described in the first section, thus the importance, usability and popularity within the community were important factors.

3.2.3. How Croatian LWP reflects on the selection

The Croatian edition of the Language White Paper was also a useful source for the evaluation and comparing available language resources in various groups of potential applicability:

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
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 SD, annotated SD, dialogue SD)	3	1	3	3	4	3	4

Multimedia and multimodal data (text with AV)	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	0	0	0	0	0	0	0

Table 5. The data from the Croatian White Paper

We can consider the results of the table very useful for showing where real gaps in Croatian language resources exist.

With regard to resources such as reference corpora, parallel corpora, morphological lexicons, the situation is reasonably good for Croatian since these fundamental resources have been developed to satisfactory level in recent years. The only serious gap is Croatian Wordnet which is still at the level of basic concept sets and is below other existing wordnets for project languages. While some reference corpora of high quality and quantity exist, i.e. the Croatian National Corpus which lemmatised and MSD-tagged, large syntactically and semantically corpora annotated by experts are not available. There is also insufficient size of parallel corpora useful for training statistical machine translation systems. Discourse annotated corpora do not exist, and the scores for multimedia and multimodal data are very low. Also, the general usage formal grammar models do not exist, apart from local grammars in NERC system and Croatian Language Processing Tools in NooJ.

3.2.4. Proportion between the selected resources for Croatian developed inside and outside the consortium

12.5% of the selected resources and tools were developed outside the consortium, 87.5% of them were developed inside the consortium, as indicated below:

Resource name	Developed inside/outside the consortium?
Croatian National corpus	internal
Croatian-English Parallel Corpus	internal
Southeast European Parallel Corpus	internal
Croatian Dependency Treebank	internal
Croatian Morphological Lexicon	internal
Croatian wordnet	internal
Croatian Valency Dictionary	internal
Croatian Language Corpus	external

Table 6. Croatian resources developed inside and outside the consortium

3.2.5. Analysis of the set of the already selected Croatian resources and tools

Regarding its technological impact, CESAR targets specific Croatian language resources with a view to improving their availability, interoperability and representativeness. The following general categories of resources strictly follow this requirement:

- written corpora and lexical databases: basic prerequisites for most NLP solutions
- language processing tools from tokenization to higher level processing

Written corpora

Manually annotated corpora are important resources, used for training and testing various language processing tools. They also provide a reference to measure the performance of tools with the same function, e.g. lemmatizers and PoS/MSD-taggers. That is the reason why Croatian monolingual corpora (e.g. Croatian National Corpus) were selected to cleanup data, to provide annotation in widely used formalisms and to improve documentation.

Parallel corpora are among the most important resources used in multilingual language processing. They serve as training data sets in machine translation systems, cross-linguistic information retrieval and in the construction of bilingual dictionaries. That is why Croatian-English Parallel Corpus was selected into the core set of resources.

Lexical databases

Parallel to mono- and/or multi-lingual corpora, the digital lexicons represent the second major type of language resources. They represent the basic prerequisite for building the rule-based systems that depend on information stored in lexica, starting with morphological information up to semantic roles information. For that reason, important lexical resources, e.g. Croatian Morphological Lexicon and Croatian Valency Dictionary have been selected.

Language processing tools

The standard preprocessing steps (tokenization, sentence segmentation, lemmatisation, PoS and full MSD tagging, including disambiguation) could be considered near completed for Croatian. These tools were selected to be upgraded and made accessible through META-SHARE platform, mainly because they are interoperable with each other. The aim is linking the processing components into a tool chain (pipeline) with a standardized API (most suitably in the form of web services), which will improve their adaptability and maturity, but provide to the users the most updated version. These tools will be standardized according to the META-NET reccomendations, and their documentation will be improved and extended.

3.2.6. Croatian resources of greatest interest in the next rounds of selection

The development of the speech technology for Croatian is still in its early stage and there are only few basic resources collected so far. Croatian spoken corpora should be recognised, documented, described and made available through META-SHARE platform. So far there are only two research groups in Croatia working on these resources and they have been contacted.

Very large parallel corpora with Croatian as one of the languages in the pair are much needed for different purposes. Upgrading the existing collection of Croatian translations of Acquis Communautaire could lead to the upgrading of the JRC Acquis Corpus with the Croatian side. This outcome would be more than welcome because it can lead to immediately applicable multilingual resource that can be used by the translation services of EC and other EU bodies.

Also, development of chunking and parsing procedures should be put forward in order to enable automatisation of annotation of large corpora at higher linguistic levels, i.e. syntactic (incl. both, shallow and deep parsing, preferably following many different grammar

formalisms) and semantic (incl. both, lexical semantic annotation with wordnet sense annotation and WSD, and sentence semantic annotation, i.e. semantic roles recognition).

3.3. Hungarian

3.3.1. General evaluation of Hungarian resources

For upgraded resources:

- **All selected resources are state-of-the-art representatives of their type for Hungarian:**
  - Hunglish parallel corpus: the largest English-Hungarian parallel corpus
  - Hungarian wordnet: the largest wordnet for Hungarian
  - Hungarian National Corpus: balanced reference corpus for Hungarian covering language variants from also beyond the border of Hungary
  - Szeged NER corpus: manually Named Entity-annotated corpus for Hungarian, which has been used to train and test named entity recognizer systems for Hungarian
  - Szeged corpus: the largest manually morphologically annotated corpus for Hungarian
  - Szeged treebank: the largest syntactically annotated corpus for Hungarian
  - Hungarian webcorpus: the largest Hungarian corpus
  - Word level speech database: words for for text-to-speech synthesis
  - Read speech database for TTS: basic units for text-to-speech synthesis
  - Named entity lexical database: pronunciation dictionary for proper nouns
  - Spoken elderly database for ASR: a semi-spontaneous speech database of Holocaust survivors
  - Lecture speech database for ASR: large Hungarian parliamentary speech corpus
  - BABEL Hungarian Clear Speech Database: a phonetically well balanced speech database suitable for research purposes, including also pronunciation modelling, prosody modelling, etc.
  - MRBA Hungarian Reference Speech Database: a phonetically well balanced speech database for text independent speech recognizers
  - MTBA Hungarian Telephone Speech Database: phoneme level segmented and annotated database, which can serve as a training database for phoneme based recognisers
  - MTÜBA Hungarian Telephone Client Speech Database: phrase level segmented spontaneous telephone conversations
  - Broadcast News Database: the largest Hungarian TV program speech database
  - Emotion database: the largest Hungarian audio emotion database
  - Sound Gesture Database: the only Hungarian spoken lexicon containing sound gestures
- **Equally valuable representatives are all included in the selection:**
  - Both rule-based and statistical processing tools are included: Hungarian Language Processing Tools in NooJ and Huntools, respectively
  - Szeged corpus and Hungarian webcorpus: both are used to train and test statistical algorithm-based NLP tools
  - Both of hunner and Named entity lexical database contains gazetteer lists
  - All of the Hungarian speech databases are used to train and test ASR and TTS systems
  - The BABEL speech database is suitable for phonetic and phonologic research, the Emotion Database (subpart of MTÜBA) is annotated for emotions

- **Current status of resources present superior quality at least on regional level without the need of excessive further development:**
  - All resources are the best available resources for Hungarian
  - All are ready to be upgraded without any development (except some audio databases)
  - **Licensing issues allow to freely process and make available the resources and resource-related materials or the consortium succeeds in reaching an agreement with respective copyright holders:**
  - All resources are either GPL-compatible, or are intended to be made available under one of the free-access licenses throughout the project (except some audio databases)
  - For extended/linked resources:
- **The extension of resources provides considerable value to the community, at least on regional level**
  - Hunglish parallel corpus: increasing the source data significantly improves its range of usage
  - Hungarian wordnet: extension the number of synsets improves its range of application
  - Hungarian National Corpus and Hungarian webcorpus: adding higher level annotation makes them applicable for wide range of linguistic analysis
  - Szeged NER corpus and Hungarian Language Processing Tools in NooJ: extension to new domains increases their coverage and adaptability
  - huntoken, hunmorph, hunpos, hunner, hunpars: linking the processing components into a tool chain with a standardized API improves their adaptability and maturity
  - Word level speech database and Read speech database for TTS: extension makes them available for research and education
  - Named entity lexical database: after extension it will be applicable to derive a scalable text-to-sound converter
  - Spoken number database for TTS: after extending the system, its model will be freely available for research and education
  - Spoken elderly database for ASR and Lecture speech database for ASR: extension the level of labelling improves their coverage
  - Medical database, Sound Gesture database and Emotion database: enlargements with new entries improves their coverage and adaptability
- **The emphasis is on providing building blocks to the existing tools rather than major restructuring**:
  - All of the planned actions are additions of synsets, source data, higher level annotation, new entries etc., not restructuring
- **Additional resources are integrated with existing ones only if they significantly improve the quality of resulting resources**:
  - No additional resources
- **If more than one representative of certain tool type for a language has been selected, they are very likely to be interlinked to benefit from strong points of both solutions**:
  - Speech databases will be interlinked to make them easier to use both in research and education
- **If less-developed, but still very popular tools can benefit from the enhancement basing on their well-developed equivalent, their enhancement is also considered**:
  - No tools involved

- **Experience of other consortium members/other consortia is extensively used in the process of further extending national resources to provide strong foundation for cross-linguality**:
  - Selection of resources and tools are based on consultation with partners both inside and outside the consortium
- **Tools offering language-neutrality or cross-linguality are preferred**:
  - Most of the tools contain language-independent modules (e.g. hunpos needs a model which can be generated from corpora in any languages)
  - Several resources follow international standards in annotation (e.g. Szeged NER corpus applies the standard CoNLL tagset)
  - For resources aligned across languages:
- **No more than one tool of a certain type for each language is used**:
  - Hunglish parallel corpus: providing cross-language alignment for the project's languages for common corpora
  - Hungarian Wordnet: aligning synsets across similar resources within the project (plWordNet, Serbian Wordnet, Bulgarian Wordnet, Croatian Wordnet)
  - Hungarian Language Processing Tools in NooJ: addition modules for project's languages
  - hunpos and hunner: extension of language model to work with other project languages
  - huntoken: extension of tokenization model to work with other project languages
  - hunmorph: building morphological descriptions to cover other project languages
- **Whenever applicable, the largest set of languages is selected**:
  - All potentially available parallel resources were gathered
- **Language-independence is targeted to a great extent**:
  - Language-independence is obtained on the representation format level
  - The nature of data for parallel corpora is obviously language-dependent
- **The quality of a result is of immense concern**:
  - The quality of the result is maintained by compliance with agreed standards on all levels of annotation
  - Special effort has been made to make the parallel corpora available not only with the research community formats (TEI), but also in an industry-wide format (XLiFF)

3.3.2. Total Point Value for Hungarian resources

The selection process was not based on TPV factors. Nevertheless, TPVs were calculated for the resources and tools being evaluated according to measures specified:

Resource	Available for whom?	For how much?	Adapta-bility	Comp-liance	Soundness	Task-relevance	Environment-relevance	Quantity	TPV
Medical database	3	1	2	3	2	2	3	3	**19**
Named entity lexical database	3	4	2	2	2	1	2	2	**18**
Hungarian National Corpus	3	1	3	1	1	2	2	3	**16**
Read speech database for TTS	3	4	2	2	1	1	1	2	**16**
BABEL Hungarian Clear SD	1	4	2	2	1	1	2	2	**15**

MTÜBA HU Telephone Client SD	2	3	2	1	1	2	2	1	**14**
Sound Gesture Database	2	1	2	1	2	1	3	2	**14**
Emotion Database	1	1	2	2	2	2	2	2	**14**
Hungarian Wordnet	1	1	3	1	1	2	3	1	**13**
Word level speech database	1	2	1	2	2	1	3	1	**13**
MRBA HUReference SD	1	4	2	1	1	1	2	1	**13**
MTBA HU Telephone SD	1	4	2	1	1	1	2	1	**13**
HU Language Processing Tools in NooJ	3	1	2	2	1	2	2	0	**13**
Hunglish parallel corpus	1	1	1	1	3	2	2	1	**12**
Szeged NER corpus	1	1	1	1	1	2	2	3	**12**
Hungarian webcorpus	1	1	1	3	2	1	2	1	**12**
hunpars	1	1	1	3	2	3	1	0	**12**
Lecture speech database for ASR	1	1	1	2	1	2	2	1	**11**
Broadcast News Database	1	1	1	1	1	1	2	3	**11**
hunner	1	1	1	1	3	3	1	0	**11**
hunmorph	1	1	1	3	2	1	1	0	**10**
Szeged corpus	1	1	1	1	1	1	2	1	**9**
Szeged treebank	1	1	1	1	1	1	2	1	**9**
Spoken elderly database for ASR	2	1	1	1	1	1	1	1	**9**
hunpos	1	1	1	2	1	1	1	0	**8**
huntoken	1	1	1	1	1	1	1	0	**7**

hunalign	1	1	1	1	1	1	1	0	**7**

Table 7. TPVs for Hungarian resources

As the table shows, resources and tools with high TPV were selected (bellow or equal to 16), supplemented with the most widely used reference corpora and the hun* tools (only two resources with bigger TPV were selected: Medical database and Named Entity lexical database). As mentioned above the selection was based on the criteria described in the first section, thus the importance, usability and popularity within the community were important factors, irrespectively of TPVs.

3.3.3. How Hungarian LWP reflects on the selection

The Hungarian edition of the Language Whitepaper was also a useful source for the evaluation and comparing available language resources in various groups of potential applicability:

	Quantity	Availability	Quality	Coverage	Maturity	Sustaina-bility	Adaptability
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 SD, labelled/annotated	2	2	4	2	4	4	0
Multimedia and multimodal data (text data with AV)	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	2	6	1	1	1	4	2

Table 8. Data from the Hungarian White Paper

We can consider the results of the table useful for showing us where real gaps in Hungarian language resources can be detected. There are no discourse corpora in Hungarian, and the scores for multimedia and multimodal data are near to zero, as well. As the main aim of the CESAR project is to upgrade and extend the existing resources, we will not build new corpora, rather focus on updating the existing ones with quite low scores (e.g. Sound gesture database for multimodal data). As it can be seen, there are not sufficient number of syntax corpora, speech corpora, wordnets and ontological resources. That is why we selected Szeged treebank, Hungarian wordnet and quite large number of speech databases for upgrading. The resources with lower scores are equally well suited for further processing since low numbers reflect the need of improvement.

3.3.4. Proportion between the selected resources for Hungarian developed inside and outside the consortium

37% of the selected resources and tools were developed outside the consortium, 40% of them were developed inside the consortium, 15% of them were developed partially internal and external, and the remaining 8% are still under negotiation, as indicated below:

Resource name	Developed inside/outside the consortium?
Hunglish parallel corpus	internal/external
Hungarian Wordnet	internal/external
Hungarian National Corpus	Internal
Szeged NER corpus	External
Szeged corpus	External
Szeged treebank	External
Hungarian webcorpus	External
Hungarian Language Processing Tools in NooJ	internal
hunner	external
huntoken	external
hunpos	external
hunmorph	external
hunpars	external
hunalign	external
Word level speech database	internal
Read speech database for TTS	internal
Named entity lexical database	internal
Spoken elderly database for ASR	under negotiation
Lecture speech database for ASR	under negotiation
BABEL Hungarian Clear Speech Database	internal
MRBA Hungarian Reference Speech Database	internal/external
MTBA Hungarian Telephone Speech Database	internal/external
MTÜBA Hungarian Telephone Client Speech Database	internal
Broadcast News Database	internal
Emotion database	internal

Sound gesture database	internal
Medical database	internal

Table. 9. Hungarian resources developed inside and outside the consortium

3.3.5. Analysis of the set of the already selected Hungarian resources and tools

Regarding its technological impact, CESAR targets specific Hungarian language processing resources with a view to improving their availability, interoperability and representativeness. The following general categories of resources strictly follow this requirement:

- written corpora and speech databases: basic prerequisites for most NLP solutions
- language processing tools from tokenization to higher level parsing

Written corpora

Manually annotated corpora are important resources, used for training and testing various language processing tools. They also provide a reference to measure the performance of tools with the same function, e.g. morphological analyzers. That is the reason why Hungarian monolingual corpora (e.g. Szeged corpus and treebank, Hungarian National Corpus) were selected to cleanup data, to provide annotation in widely used formalisms and to improve documentation.

Parallel corpora are among the most important resources used in multilingual language processing. They serve as training data sets in machine translation systems, cross-linguistic information retrieval and in the construction of bilingual dictionaries. That is why Hunglish parallel corpus was selected into the core set of resources.

Speech databases

Representative and phonetically balanced corpora are the most important resources, including annotation (noise events, mispronunciations, phonetic transcription/lexicon). Resources ensuring the exploration of speech prosody, emotions in speech or voice pathology are also of interest.

Language processing tools

The standard preprocessing steps (tokenization, morphological analysis, POS tagging) are completed for Hungarian, moreover there are three morphological analyzers for Hungarian. The hun* tools were selected to upgrade and upload to META-SHARE repository, mainly because they are interoperable with each other. The aim is linking the processing components into a tool chain with a standardized API, which will improve their adaptability and maturity. The hun* tools will be standardised according to the META-NET rules, and their documentation will be improved and extended.

3.3.6. Hungarian resources of greatest interest in the next rounds of selection

Speech technology is shifting towards speech understanding, dialogue modelling, multimodality, hence in the future databases containing parallel audio-video would be of high interest. Annotation should contain dialogue act annotation, prosodic annotation, emotions etc. in order to allow for supporting modern speech technology claims. A spoken corpora recorded in car environment would be also useful, however, as far as we know no such project is currently running that would plan the recording of such a resource.

3.4. Polish

3.4.1. General evaluation of Polish resources

For upgraded resources:

- **All selected resources are state-of-the-art representatives of their type for Polish:**
  - NKJP – the largest monolingual corpus of Polish.
  - The corpus of frequency dictionary of Polish language of the 1960s – the source of training data for taggers.
  - Morfeusz, Morfologik – two largest morphological dictionaries of Polish.
  - plWordNet – the largest wordnet of Polish.
  - Polish Treebank – the largest and most detailed treebank of Polish.
  - Polish Parallel Corpora – the set of practically all potentially available Polish parallel corpora.
  - Polish Spoken Multimedia Corpus – first spoken multimedia resource for Polish.
  - Polish Sejm Corpus – large publicly available dataset with partial audio/video coverage.
  - NE Resources with Gazetteers – largest set of Polish NE resources.
  - Polish Causal Spoken Discourse Corpus – the largest casual spoken discourse corpus of Polish.
  - Valency dictionary – is created after merging 3 other dictionaries – the milestones in Polish valency research.
- E**qually valuable representatives are all included in the selection:**
  - Both Morfeusz and Morfologik are included.
  - Polish Parallel Corpora: practically all available parallel corpora of Polish are gathered for the selection.
  - Polish Spoken Multimedia Corpus – all spoken resources are included.
  - NE Resources with Gazetteers – all Polish NE resources are included.

- **Current status of resources present superior quality at least on regional level without the need of excessive further development:**
  - All resources are best available resources for Polish.
  - All are ready to be upgraded without any development.

- **Licencing issues allow to freely process and make available the resources and resource-related materials or the consortium succeeds in reaching an agreement with respective copyright holders:**
  - All resources are either GPL-compatible or are intended to be made available under one of the free-access licenses throughout the project.
  - For extended/linked resources:

For extended/linked resources:

- **The extension of resources provides considerable value to the community, at least on regional level:**
  - The corpus of frequency dictionary of Polish language of the 1960s – manual tagging of the resource is planned to improve the amount of all available language data to be used for Polish tagger training by 50%.
  - Morfeusz, Morfologik – merger of both resources and extension of the resulting single resource is going to create the largest morphological dictionary of Polish.

- Polish Spoken Multimedia Corpus – annotation of speech acts creates the first resource of such size for Polish.
- NE Resources with Gazetteers – merger and development of Polish NE resources creates considerable value for the Polish language community.
- Casual Spoken Discourse Corpus – addition of time alignment annotation for a subset of the corpus significantly improves its range of applications.
- Parallel corpora are enriched with bibliographic and structural metadata, which improves their applicability.
- Polish Valency Dictionary – merger of existing resources and extension of the resulting single resource is going to create the largest valency dictionary of Polish.
- Cross-lingual Repository of Named Entities – cross-lingual merger of Polish and multilingual NE dictionaries.

- **The emphasis is on providing building blocks to the existing tools rather than major restructuring:**
  - Mergers seem like major restructuring, but they are necessary to create the best quality since e.g. morphological dictionaries are using slightly different taggers and existing valency dictionaries – incompatible notations. All mergers are based on the building blocks provided by components..
  - Tag converters are planned to be created in the course of merging the dictionary bases and can be further reused to process other available morphological data.

- **Additional resources are integrated with existing ones only if they significantly improve the quality of resulting resources:**
  - Morfeusz, Morfologik – the value of the largest morphological dictionary of Polish is indisputable; its quality will be ensured by manual verification of records and will be constantly assessed by linguists and the community.
  - NE Resources with Gazetteers – the merger creates the largest NE resource for Polish; its quality will be ensured by manual verification of entries.
  - Polish Valency Dictionary – similarly to the morphological case, creating the largest valency dictionary of Polish provides a real value to the community.

- **If more than one representative of certain tool type for a language has been selected, they arevery likely to be interlinked to benefit from strong points of both solutions:**
  - To improve quality of the resulting resources, the decision was made to create deep mergers of morphological dictionaries, valency dictionaries and NE resources rather than simple interlinking.

- **If less-developed, but still very popular tools can benefit from the enhancement basing on their well-developed equivalent, their enhancement is also considered:**
  - This requirement is valid for The corpus of frequency dictionary of Polish language of the 1960s – although the resource seems old, it is still very popular due to its relatively large size and manual verification.
  - Further manual improvement and upgrade to NKJP format is considered valuable.

- **Experience of other consortium members/other consortia is extensively used in the process of further extendinnational resources to provide strong foundation for cross-linguality:**

- Selection of formats, e.g. for parallel corpora, was based on inter-consortium agreement.

- **Tools offering language-neutrality or cross-linguality are preferred:**
  - The branch of parallel corpora was selected as one of the most valuable resources resulting from the project due to their cross-lingual nature.
  - While other Polish resources are compatible with common requirements for language standards, they are still underdeveloped as compared to other languages and this requirement was considered less important.

For resources aligned across languages:

- **No more than one tool of a certain type for each language is used:**
  - The single set of parallel corpora and NE dictionaries is processed.
  - Possible overlaps within the consortium are eliminated.
- **Whenever applicable, the largest set of languages is selected:**
  - All potentially available parallel resources were gathered.
- **Language-independence is targeted to a great extent:**
  - Language-independence is obtained on the representation format level.
  - The nature of data for parallel corpora is obviously language-dependent.
- **The quality of a result is of immense concern:**
  - The quality of the result is maintained by compliance with agreed standards on all levels of annotation.
  - Special effort has been made to make the parallel corpora available not only with the research community formats (TEI), but also in an industry-wide format (XLiFF).

3.4.2. Total Point Value for Polish resources

The TPV factors were calculated for the resources being evaluated according to measures specified:

Resource	Available for whom?	Available for how much?	Adapta-bility	Compliance	Sound-ness	Task-relevance	Environment-relevance	Quantity	TPV
Polish Treebank	3	1	1	3	1	1	2	1	**13**
PL Parallel Corpora	2	2	2	2	2	1	1	1	**13**
Polish Sejm Corpus	1	1	1	2	2	2	2	2	**13**
Corpus of frequency dictionary	1	1	1	2	1	1	2	3	**12**
PL Spoken Multimedia Corpus	2	1	1	2	1	2	1	2	**12**
PL Spoken Discourse Corpus	2	2	2	2	1	1	1	1	**12**
NE with Gazetteers	1	1	1	2	1	1	2	2	**11**

Valence dictionaries	1	1	1	2	1	1	2	2	**11**
Morphological dictionary Morfeusz	2	1	2	1	2	1	1	1	**10**
Morphological dictionary Morfologik	1	1	1	1	3	1	1	1	**9**
Formal Grammar of Polish	1	1	1	2	1	1	1	1	**9**
IPI PAN Corpus	1	1	1	1	1	1	1	2	**9**
Shallow grammar for Spejd	1	1	1	1	1	1	1	2	**9**
Polish Internet Corpus	1	1	1	1	1	1	1	2	**9**
Polish collocation dictionary	1	1	1	1	1	1	1	1	**8**
plWordNet	1	1	1	1	1	1	1	1	**8**
National Corpus of Polish	1	1	1	1	1	1	1	1	**8**
Polish Valency Dictionary	3	1	1	1	2	1	1	1	**11**
Cross-lingual Repository of Named Entities	3	1	1	1	1	1	1	1	**10**

Table 10: TPVs for Polish resources

Best-suited resources were selected and supplemented (all of the with TPV bellow 16) with two resources of the greatest importance: the Polish Wordnet (project in progress) and NKJP – the largest monolingual corpus of Polish.

3.4.3. How Polish LWP reflects on the selection

The Polish edition of the Language White Paper is also a useful source for the evaluation and comparing available language resources in various groups of potential applicability:

	Quantity	Availability	Quality	Coverage	Maturity	Sustaina-bility	Adaptability
Reference Corpora	3	1	4	4	5	5	5
Syntax-Corpora (treebanks, dependency banks)	3	2	4	4	5	5	3

Semantics-Corpora	3	4	4	4	3	4	4
Discourse-Corpora	3	2	4	4	4	4	2
Parallel Corpora, Translation Memories	2	1	4	4	1	4	4
Speech-Corpora (raw SD, labelled/annotated SD, etc.)	3	2	4	4	3	2	2
Multimedia and multimodal data (text data with AV)	1	0	4	2	2	0	4
Language Models	1	0	3	3	2	2	2
Lexicons, Terminologies	2	1	4	2	1	1	2
Grammars	1	3	1	1	1	1	1
Thesauri, WordNets	1	1	2	1	1	1	1
Ontological Resources for World Knowledge	1	1	1	1	1	0	0

Table 11. Data from the Polish White Paper

The evaluation of Polish multimedia and spoken corpora is very low with respect to their availability and adaptability, which justifies our choice of the Multimedia Speech and Conversational Spoken Polish Corpus for the inclusion in the set of core Polish CESAR resources.

3.4.4. Proportion between the selected resources for Polish developed inside and outside the consortium

Less than 25% of resources selected for processing were developed inside the consortium while 30% were developed outside the consortium and over 45% of resources were partially internal and external, as indicated below:

Resource name	Developed inside/outside the consortium?
NKJP	internal/external
Korpus słownika frekwencyjnego	external
Morfologik	external
Morfeusz	internal
plWordNet	external
Polish Treebank	internal

Polish Parallel Corpora	external
Polish Spoken Multimedia Corpus	internal/external
Polish Sejm Corpus	internal/external
NE Resources with Gazetteers	internal/external
Polish Casual Spoken Discourse Corpus	internal

Table 12. Polish resources developed inside and outside the consortium

3.4.5. Analysis of the set of the already selected Polish resources and tools

Regarding its technological impact, CESAR targets specific Polish language processing resources with a view to improving their *availability*, *interoperability* and *representativeness*. The following general categories of resources strictly follow this requirement:

- Morphological dictionaries and annotated corpora – a basic prerequisite for most NLP solutions.
- Spoken discourse corpora and speech databases – sparsely distributed across different languages.
- New and existing parallel corpora annotated in widely accepted text encoding and translation memory formats.
- Higher-level syntactic and semantic resources such as Wordnets, dictionaries of named entities, valency dictionaries and treebanks of Polish.

Dictionaries

Morphological dictionaries are about the most basic language resources, and most NLP tasks require their existence and availability. Until recently, there has only been one morphological dictionary available for Polish under an open source licence (LGPL and Creative Commons), namely, Morfologik (http://morfologik.blogspot.com/ ; not to be confused with the Hungarian NLP company Morphologic). Another morphological analyser, Morfeusz (http://sgjp.pl/morfeusz/), whose quality is widely believed to be higher than that of Morfologik, was available under a closed – albeit free for non-commercial applications – licence. These two tools seemed to be the most widely used morphological analysers for Polish; actually, both were used in the National Corpus of Polish (http://nkjp.pl/).

CESAR intended to obtain the agreement of the owners of the data of both dictionaries to release them on a very liberal open source licence (the FreeBSD licence, also known as the 2-clause BSD licence). Moreover, by initiating the cooperation between the maintainers of the dictionaries, a single large morphological dictionary for Polish was planned to be created, comprising and extending both Morfologik and Morfeusz.

Annotated Corpora

Manually annotated corpora are important resources, used for training various language processing tools. One of the most basic such tools are morphological taggers, used for disambiguating the results of morphological analysers. The most comprehensive resource of this kind for Polish is the 1-million-word subcorpus of the National Corpus of Polish (PL: Narodowy Korpus Języka Polskiego; NKJP), manually annotated at various linguistic levels, including the morphosyntactic level. However, for a morphologically rich language, 1 million words is not sufficient to attain the same tagging accuracy as, for example, for English (over 97%); in fact, current Polish taggers perform at the level of 92–93%.

In order to improve these results, two kinds of activities were undertaken in CESAR. First, although a very careful annotation procedure was adopted in NKJP, annotation errors may

readily be found in the corpus, so known issues are corrected manually and semi-automatically within CESAR. Additionally, statistical methods are employed to discover unknown errors.

Second, an additional corpus of 500 thousand words is annotated within CESAR, with the aim of creating a high-quality 1.5-million-word training corpus. However, in order to minimise costs, an existing corpus is used for this purpose, namely, the "Polish language of the 1960s" corpus (http://clip.ipipan.waw.pl/PL196x). The corpus was originally manually annotated with a much more limited tagset than that currently used for Polish, so the work consists in the semi-automatic conversion the annotation of that corpus to the current standards and – most importantly – in its independent re-annotation. These two annotations are compared and any differences are sent for adjudication, thus increasing the annotation quality.

Spoken Corpora

Corpora of casual spoken discourse are a rather rare resource for many languages. 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 will be made available in the TEI P5 format following some privacy considerations. Furthermore, a selection of the transcriptions are being time-aligned with the original recordings at the level of utterances and made available under the GPL license through the META-SHARE repository. Another multimedia speech corpus planned to be included into META-SHARE repository is the TEI-encoded corpus of transliterated complex spontaneous human-human telephone conversations acquired in the course of LUNA (Spoken Language UNderstanding in multilinguAl communication systems; http://www.ist-luna.eu) project. The source data have been collected at the call centre of the Public Transport Authority of Warsaw and annotated in terms of semantic constituents and semantic structures.

Parallel Corpora

Parallel corpora are among the most important resources used in multilingual language processing. On the one hand, they serve as training data sets in machine translation systems, cross-linguistic information retrieval and in the construction of bilingual dictionaries. Depending on their annotation format, they can also be used, more or less readily, as translation memories, as well as an empirical basis in comparative linguistic and translation studies. 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. First of all, they are available from a relatively large number of different sources, which often makes it difficult to identify the right set of corpora to use for a particular purpose. Secondly, when it comes to annotation standards, Polish parallel corpora and translation memories come in many shapes and sizes. Some resources are available as translation memories without any text structure annotation. Some parallel corpora are little more than plain text collections which encode segment boundaries using simple line breaks, while others make use of sophisticated annotation schemas which make it possible to express non-trivial cases of equivalence between segments, such as non-sequential cross-links,

deletions, insertions, or segment splits and mergers. Apart from technical problems, the representativeness of openly available parallel resources leaves much to be desired. The majority of freely available parallel corpora and translation memories are public domain legal collections and open license software and technical documentation localization memories. This in turn means that non-technical and non-legal text genres and language registers tend to be poorly represented in openly available corpora.

The first contribution of the CESAR project to the availability, interoperability and representativeness of parallel corpora of Polish is a source containing some 500 scientific articles in Polish and English from Academia – the Magazine of the Polish Academy of Sciences. The articles were first converted from the PDF format and aligned semi-automatically at the sentence level using the memoQ CAT environment. The initial sentence-level alignment was then manually verified and the texts were further annotated with bibliographic information. Since this data is a completely new parallel resource for Polish, it can be considered as an example of CESAR's contribution to improving the coverage and representativeness of Polish parallel corpora. The CORDIS collection contains over 10 000 articles published at http://cordis.europa.eu/ – the Community Research and Development Information Centre in Polish and 5 other EU languages. The CORDIS and RAPID collections were web-crawled parsed for contents and aligned. Compared with the 2010 version of these two collections available from the Information Systems Laboratory at the Adam Mickiewicz University, the CESAR version features structural and bibliographic annotation adhering to the TEI and XLiFF formats described below. We have also decided to include the Polish-English component of the JRC version of Acquis Communautaire in the first batch of resources for the sake of its increased availability in the output formats. In the process of converting the JRC Acquis corpus additional bibliographic information was added to the metadata headers.

Once the variously encoded collections are converted and normalised, they can be processed and exported into more uniform and standard formats used for the exchange of parallel corpora and translation memories. We have decided to provide the parallel data in two main formats TEI and XLiFF. The first format is a widely recognised standard of annotating corpus data with good support for encoding structural, bibliographic and alignment annotation. We expect this format to be a more natural choice in corpus analysis and NLP contexts as it can be used to mark up information about alignment splits, mergers as well as many-to-many segment linking, which proves useful when manually aligning texts.

The XLiFF format, on the other hand, although much less expressive, is supported by all major CAT environments as an increasingly popular way of exchanging translation memories. Any subset of the parallel collections can thus be used directly as a translation memory in a modern CAT environment.

Other resources

Apart from the above-mentioned core resources, the processing of which seems the most time-consuming and labour-intensive, another set of equally important resources will be made available through META-SHARE channels. The most prominent of them is the Polish Wordnet (http://www.plwordnet.pwr.wroc.pl), still actively developed and therefore planned to be issued in all three CESAR batch editions.

Another important resource is the merger of existing dictionaries of Polish Named Entities. Various resources are planned to be gathered (e.g. from Tours, Poznan, Warszawa and Wrocław) and standardised within this task by encoding them in the LMF (Lexical Markup Framework; ISO/IS 24613 2008) format.

Last but not least, Marcin Woliński's treebank of Polish constructed using automatic syntactic analysis will be made available in the second batch.

3.4.6. Polish resources of greatest interest in the next rounds of selection

Parallel corpora have been given one of the highest priorities in the upgrade/extension/alignment process, so more resources of this type are intended to be included in the next batches of resources (to be released in June 2012 and January 2013) or even going beyond the end of the project. A number of complementary collections of parallel corpora of Polish such as OpenTran, Opus have already been identified.

3.5. Serbian

3.5.1. General evaluation of Serbian resources

For upgraded resources:

- **All selected resources are state-of-the-art representatives of their type for Serbian:**
  - SrpKor – Corpus of Contemporary Serbian, the largest monolingual corpus of Serbian.
  - SrpLemKor - Serbian MSD Annotated Corpus, the source of training data for taggers.
  - SrpRec - Serbian Morphological Dictionary, the largest morphological dictionary of Serbian.
  - SrpWN – the largest wordnet of Serbian.
  - SrpEngKor - English-Serbian Aligned Corpus, predominantly non-fiction.
  - SrpFranKor - French-Serbian Aligned Corpus, predominantly fiction.
  - Verne80days - Multilingual Edition of Verne's Novel "Around the World in 80 Days", the largest multi-lingual text that includes Serbian.
  - SrpNER - Serbian Named Entity Resources, largest set of Serbian NE resources.
  - CSL - Corpus of Serbian Language, largest diachronic corpus of Serbian.
  - DABI - Digital Archive of the Institute for Balkan Studies, the largest multi-modal database.
  - AlfaNum MD - AlfaNum Morphologic Dictionary of Serbian, morphological dictionary with accentuation.
  - AlfaNumKor - AlfaNum Text Corpus of Serbian, largest accentuated and PoS tagged corpus of Serbian.
  - AlfaNum ASR – AlfaNum Speech Databases for ASR, largest speech corpus for speech recognition.
  - AlfaNum TTS - AlfaNum Speech Databases for ASR, largest speech corpus for text-to-speech synthesis.
- **Equally valuable representatives are all included in the selection:**
  - Both SrpRec and AlphaNum are included, they represent overlapping resources that cannot be merged.
  - Both SrpLemKor and AlfaNumKor are included, they represent overlapping resources that cannot be merged.
  - Both SrpKor and CSL are included; they represent overlapping resources that cannot be merged.
- **Excessive further development:**
  - All resources are best available resources for Serbian.
  - The most of them are ready to be upgraded without any development.
- **Licensing issues allow to freely process and make available the resources and resource-related materials or the consortium succeeds in reaching an agreement with respective copyright holders:**
  - The most of resources are either GPL-compatible, or are intended to be made available under one of the free-access licenses throughout the project.
  - The exceptions are the diachronic corpus and speech resources for which licenses will be negotiated throughout the project.

For extended/linked resources:

- **The extension of resources provides considerable value to the community, at least on regional level:**

- Serbian aligned corpora (SrpEngKor, SrpFranKor, Verne80days): increasing the source data significantly improves its range of usage while enrichment with bibliographic and structural metadata improves their applicability.
- Serbian Wordnet: extension the number of synsets improves its range of application,
- Corpus of Contemporary Serbian: adding higher level annotation (functional styles, registers, etc.) makes them applicable for wide range of linguistic analysis.
- Serbian NE Resources – extension with new NEs, normalization of Nes, etc. increases their coverage and adaptability
- Serbian MSD Annotated Corpus (SrpLemKor), experiments with various taggers, addition of new tags, enhancement with new texts will significantly improves its range of usage.
- the Serbian Morphological Dictionary (SrpRec) – extension with new entries, particularly multi-word units, and addition of new semantic markers will increase its coverage and applicability.

- **The emphasis is on providing building blocks to the existing tools rather than major restructuring:**
  - All of the planned actions are additions of synsets, source data, higher level annotation, new entries etc., not restructuring.
- **Additional resources are integrated with existing ones only if they significantly improve the quality of resulting resources:**
  - No additional resources.
- **If more than one representative of certain tool type for a language has been selected, they are very likely to be interlinked to benefit from strong points of both solutions:**
  - Similar resources (corpora and morphological dictionaries) they cannot be merged or linked because they were produced using different methodology and because of different licensing status.
- **If less-developed, but still very popular tools can benefit from the enhancement basing on their well-developed equivalent, their enhancement is also considered:**
  - No tools involved.
- **Experience of other consortium members/other consortia is extensively used in the process of further extending national resources to provide strong foundation for cross-linguality:**
  - Selection of resources and tools are based on consultation with partners both inside and outside the consortium.
- **Tools offering language-neutrality or cross-linguality are preferred**:
  - Several resources follow international or professional standards in annotation (morphological dictionaries - LADL format, MSD tagged corpus - MULTEXT-East, aligned corpora – TEI and TMX).
  - Most of Serbian resources are still underdeveloped as compared to other languages and this requirement was considered less important.

For resources aligned across languages

- **No more than one tool of a certain type for each language is used:**
  - Serbian aligned corpora: providing cross-language alignment for the project's languages for common corpora (most particularly for Verne80days).
  - Serbian Wordnet: aligning synsets across similar resources within the project (plWordNet, Hungarian Wordnet, Bulgarian Wordnet, Croatian Wordnet).
  - Serbian Language Processing Tools in NooJ: prepared on the basis of major Serbian textual and lexical resources.

- **Whenever applicable, the largest set of languages is selected:**
  - All potentially available parallel resources were gathered.
- **Language-independence is targeted to a great extent:**
  - Language-independence is obtained on the representation format level.
- **The quality of a result is of immense concern:**
  - The quality of the result is maintained by compliance with agreed standards on all levels of annotation.

3.5.2. Total Point Value for Serbian resources

The selection process was not based on TPV factors, but the criteria described above. The TPV factors were calculated for the resources being evaluated according to measures specified in CESAR:

Resource	Available for whom?	Available for how much?	Adapta-bility	Comp-liance	Sound-ness	Task-relevance	Environment-relevance	Quantity	**TPV**
SrpKor	3	1	2	1	1	1	1	1	**11**
SrpLemKor	1	1	1	2	2	1	1	1	**10**
SrpRec	2	1	2	1	2	1	1	1	**11**
SrpWN	1	1	1	1	1	1	1	1	**8**
SrpEngKor	2	1	2	1	2	1	1	1	**11**
SrpFranKor	2	1	2	1	2	1	1	1	**11**
Verne80days	2	1	2	1	2	1	1	1	**11**
SrpNER	3	1	2	1	2	1	1	1	**12**
CSL	3	1	2	3	1	1	2	1	**14**
DABI	3	1	1	2	2	1	1	1	**12**
AlfaNum MD	3	1	1	1	1	1	1	1	**10**
AlfaNumKor	3	1	1	1	1	1	1	1	**10**
AlfaNum ASR	3	1	1	1	1	1	1	1	**10**
AlfaNum TTS	3	1	1	1	1	1	1	1	**10**

Table 13. TPVs for Serbian resources

As the table shows, resources and tools with a relatively low TPV (bellow 16) were selected for the first batch, even for the external resources. The selection was based on the criteria described in the first section, thus the importance, usability and popularity within the community were important factors, irrespectively of TPVs.

3.5.3. How Serbian LWP reflects on the selection

The Serbian edition of the Language Whitepaper was also a useful source for the evaluation and comparing available language resources in various groups of potential applicability:

	quantity	availability	quality	coverage	maturity	sustaina-bility	adaptability
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 SD, labelled/annotated SD, etc.)	1	2	4	4	3	3	3
Multimedia and multimodal data (text data with AV)	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	1	1	0	1	0	1	1

Table 14. Data from Serbian White Paper

We can consider the results of the table useful for showing us where real gaps in Serbian language resources can be detected. There are no syntax, semantic and discourse corpora in Serbian, and the scores for grammars and ontological resources are near to zero, as well. As the main aim of the CESAR project is to upgrade and extend the existing resources, we will not build new corpora, rather focus on updating the existing ones with quite low scores (e.g. grammars, ontological resources, and multimodal data). The resources with lower scores are equally well suited for further processing since low numbers reflect the need of improvement.

3.5.4. Proportion between the selected resources for Serbian developed inside and outside the consortium

60% of resources selected for processing were developed inside the consortium while 40% were developed outside the consortium, as indicated below:

Resource name	Developed inside/outside the consortium?
SrpKor	internal
SrpLemKor	internal
SrpRec	internal
SrpWN	internal
SrpEngKor	internal
SrpFranKor	internal
Verne80days	internal
SrpNER	Internal
CSL	external
DABI	external
AlfaNum MD	external
AlfaNumKor	external
AlfaNum ASR	external
AlfaNum TTR	external

Table 15. Serbian resources developed inside and outside the consortium

3.5.5. Analysis of the set of the already selected Serbian resources and tools

Regarding its technological impact, CESAR targets specific Serbian language processing resources with a view to improving their availability, interoperability and representativeness. The following general categories of resources strictly follow this requirement:

- Morphological dictionaries.
- Written corpora and speech databases: basic prerequisites for most NLP solutions.
- Language processing tools from tokenization to higher level parsing.

Morphological dictionaries and lexical databases

Morphological dictionaries are about the most basic language resources, and most NLP tasks require their existence and availability. Serbian morphological e-dictionary consists of simple and multi-word units and through the workstation LeXimir tightly cooperates with higher-level lexical resources like Wordnets and NER resources and tools.

Written corpora

Written corpora are important resources for various language processing tools and applications. That is the reason why several Serbian monolingual corpora were selected: the Corpus of Contemporary Serbian used by many linguists and lexicographers, the Morphosyntactic and PoS tagged corpus used for training various NLP applications, the Corpus of Serbian language used for psychological and psycholinguistic research, and AlfaNum Text Corpus of Serbian used in various speech applications.

Parallel corpora are among the most important resources used in multilingual language processing. They serve as training data sets in machine translation systems, cross-linguistic information retrieval and in the construction of bilingual dictionaries. That is why several Serbian parallel corpora were selected into the core set of resources: English/Serbian, French/Serbian and multilingual Verne80days corpus.

Speech databases

Speech databases for speech recognition and text-to-speech synthesis were developed by **AlfaNum** group and were widely commercially applied.

Language processing tools

The standard preprocessing steps (tokenization, morphological analysis, POS tagging, lemmatization) are completed for Serbian and integrated in Nooj/Unitex environment.

3.5.6. Serbian resources of greatest interest in the next rounds of selection

Aligned corpora and NER system and resources have been given one of the highest priorities in the upgrade/extension/alignment process, so more resources of this type are intended to be included in the next batch of resources to be released in June 2012. Serbian morphological dictionary (including MWUs and encyclopedic knowledge) and tools that rely on them (Nooj/Unitex) have the highest priorities in the upgrade/extension/alignment process, so more resources of this type are intended to be included in the last batch of resources to be released in January 2013. Work on all these resources will go beyond the end of the project.

3.6. Slovak

3.6.1. General evaluation of Slovak resources

For upgraded resources:

- A**ll selected resources are state-of-the-art representatives of their type for Slovak:**
  - SNK – Slovak National Corpus, a representative corpus of contemporary written Slovak language (including a manually morphologically annotated subcorpus).
  - Corpus of Legal Texts – corpus consisting of Slovak Republic legislature.
  - Slovak Web Corpus – corpus of Slovak language .sk domain texts.
  - Slovak-Czech Parallel Corpus, sentence-level aligned parallel corpus.
  - Slovak-English Parallel Corpus, sentence-level aligned parallel corpus.
  - Corpus of Spoken Slovak, a corpus of modern standard spoken Slovak.
  - Slovak Language Treebank – manually syntactically annotated corpus.
  - Morphology database – a database of complete inflectional paradigms and morphological tags for 77 thousand lemmas.
  - Slovak WordNet – ontology database.
- E**qually valuable representatives are all included in the selection:**
  - Both Slovak-English and Slovak-Czech parallel corpora are included in the selection, being of comparable size and quality, whereas Slovak-Russian and Slovak-French corpora are not, because of their limited size and lower quality of linguistic annotation.

  - There is an internal corpus of Slovak texts downloaded from the internet collected at the Technical University of Košice, it is however not publicly available (neither are there any plans by the authors to make it available).
  - There is another Slovak Language morphology database developed at the Charles University, the licensing terms are restricted (commercial availability, individual agreement).
  - There is yet another Slovak Language morphology database being developed at the Masaryk University in Brno, it is however not finished, and the prospective licensing is unknown.
- **Current status of resources present superior quality at least on regional level without the need of excessive further development:**
  - All resources are best available resources for Slovak.
  - All are ready to be upgraded without any development with the exception of Slovak WordNet, which is a work in progress.
- **Licensing issues allow to freely process and make available the resources and resource-related materials or the consortium succeeds in reaching an agreement with respective copyright holders:**
  - All resources with distribution rights held by Ľ. Štúr Institute of Linguistics are available under OpenSource/OpenContent licenses (mostly triple-licensed under Affero GPL, CreativeCommons Attribution-ShareAlike and GNU Free Documentation License).
  - The written corpora are available for queries, access to the text is restricted by existing copyright law and license agreements with text providers (pseudocorpora are freely available).
  - Corpus of Spoken Slovak is not subject to copyright restrictions and is freely available.

For extended/linked resources:

- **The extension of resources provides considerable value to the community, at least on regional level:**
  - The main Slovak National Corpus is the basic reference database for the Slovak Language NLP, it is of an immense value to the community. The corpus is being continuously extended by new texts.
  - Morphology Database will be extended by additional lemmas, which allows it to increase the coverage of analysed texts.
  - Corpus of Spoken Slovak is the only public database of contemporary Slovak speech with reasonably accurate phonemic transcription, however the geographical and demographic distribution of speakers is rather unbalanced. Increasing the size will help balancing the data.
- **The emphasis is on providing building blocks to the existing tools rather than major restructuring:**
  - No major restructuring is planned.
- **Additional resources are integrated with existing ones only if they significantly improve the quality of resulting resources:**
  - No additional resources are planned.
- **If more than one representative of certain tool type for a language has been selected, they are very likely to be interlinked to benefit from strong points of both solutions:**
  - No tools are involved.

- **If less-developed, but still very popular tools can benefit from the enhancement basing on their well-developed equivalent, their enhancement is also considered:**
  - No tools are involved.
- **Experience of other consortium members/other consortia is extensively used in the process of further extending national resources to provide strong foundation for cross-linguality:**
  - Selection of resources and their format was based on inter-consortium agreement.
- **Tools offering language-neutrality or cross-linguality are preferred:**
  - The parallel corpora was selected as one of the most valuable resources resulting from the project due to their cross-lingual nature.
  - Other resources are compatible with common requirements for language standards.
  - There exists a conversion from the Slovak National Corpus morphosyntactic tagset into the Multext East tagset.

For resources aligned across languages:

- **No more than one tool of a certain type for each language is used:**
  - The single set of parallel corpora is processed, possible overlaps within the consortium are eliminated.
- **Whenever applicable, the largest set of languages is selected**:
  - All good quality potentially available parallel resources were gathered.
- **Language-independence is targeted to a great extent**:
  - Language-independence is obtained on the representation format level.
  - The nature of data for parallel corpora is obviously language-dependent.
- **The quality of a result is of immense concern**:
  - The quality of the result is maintained by compliance with agreed standards on all levels of annotation.

3.6.2. Total Point Value for Slovak resources

The selection process was not based on TPV factors, but the TPV factors were calculated for the resources being evaluated according to measures specified in CESAR:

Resource	Available for whom?	Available for how much?	Adaptability	Standard-compliance	Soundness	Task-relevance	Environment-relevance	Quantity	TPV
Slovak National Corpus	3	1	2	2	1	1	2	1	**13**
Corpus of Legal Texts	3	1	2	2	1	1	2	1	**13**
Slovak Web Corpus	1	1	2	2	2	1	2	1	**12**
Slovak-Czech PC	3	1	2	2	1	1	1	1	**12**
Slovak-English PC	3	1	2	2	1	1	1	1	**12**
Corpus of Spoken Slovak	1	1	2	2	1	1	1	1	**10**

Slovak Language Treebank	2	1	2	2	2	1	1	1	**12**
Morphology database	1	1	2	1	2	1	1	2	**11**
Slovak WordNet	1	1	2	2	2	1	1	1	**11**

Table 16. TPVs for Slovak resources

As the table shows, resources and tools with a relatively low TPV (bellow 16) were selected for the first batch. The selection was based on the criteria described in the first section, thus the importance, usability and popularity within the community were important factors, irrespectively of TPVs.

3.6.3. How Slovak LWP reflects on the selection

The Slovak edition of the Language White Paper is also a useful source for the evaluation and comparing available language resources in various groups of potential applicability:

	Quantity	Availability	Quality	Coverage	Maturity	Sustaina-bility	Adaptability
Reference Corpora	2	3	2	2	2	2	3
Syntax-Corpora (treebanks, dependency banks)	2	4	4	5	4	4	4
Semantics-Corpora	2	5	2	1	2	4	4
Discourse-Corpora	3	2	3	4	3	4	3
Parallel Corpora, Translation Memories	1	4	2	2	2	3	3
Speech-Corpora (raw SD, labelled/annotated SD, etc.)	2	3	3	2	1	2	1
Multimedia and multimodal data (text data with AV)	0	0	0	0	0	0	0
Language Models	3	4	2	2	3	3	3
Lexicons, Terminologies	0	0	0	0	0	0	0
Grammars	1	4	1	3	3	3	4
Thesauri, WordNets	1	1	2	1	3	2	3

Ontological Resources for World Knowledge	1	1	3	2	2	3	3

Table 17. Data from the Slovak White Paper

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. resources to be multilingually aligned, all factors are crucial and the resource should represent the best available result; thus the parallel corpora were selected for this group. On the other hand, the resources with lowest scores are equally well suited for further processing since low numbers reflect the need of improvement of particular language area. resources with zero scores indicate gaps in the general Slovak HLT data that badly need to be filled in – but this is outside of the scope of CESAR.

3.6.4. Proportion between the selected resources for Slovak developed inside and outside the consortium

89% of resources selected for processing were developed inside the consortium while 11% of resources (one) was partially internal and external, as indicated below:

Resource name	Developed inside/outside the consortium?
Slovak National Corpus	internal
Corpus of Legal Texts	internal
Slovak Web Corpus	external/internal
Slovak-Czech Parallel Corpus	internal
Slovak-English Parallel Corpus	internal
Corpus of Spoken Slovak	internal
Slovak Language Treebank	internal
Morphology database	internal
Slovak WordNet	internal

Table 18. Slovak resources developed inside and outside the consortium

3.6.5. Analysis of the set of the already selected Slovak resources and tools

Regarding its technological impact, CESAR targets specific Slovak language processing resources with a view to improving their availability, interoperability and representativeness. The following general categories of resources strictly follow this requirement:

- Written corpora and speech databases – a basic prerequisite for most NLP solutions.
- Language processing tools from tokenization to higher level parsing.

Written corpora

Huge, representative "national" corpora are the basic, indispensable source of important data for many NLP related tasks. They always benefit from additional data and more accurate annotation.

Manually annotated corpora are important resources, used for training and testing various language processing tools. They also provide a reference to measure the performance of tools with the same function, e.g. morphological analysers. This is why we have chosen Slovak National Corpus, Slovak Web Corpus and Corpus of Legal Texts as three independent big

monolingual corpora (each within its own domain), with manually morphologically annotated corpus being part of the Slovak National Corpus.

Parallel corpora are among the most important resources used in multilingual language processing. They serve as training data sets in machine translation systems, cross-linguistic information retrieval and in the construction of bilingual dictionaries. This was the reason of selecting two mature parallel corpora, Slovak-English and Slovak-Czech ones.

Speech databases (spoken corpora)

Spoken corpora with reliable phonemic transcription are important resources for spoken language processing and research. This was the reason for choosing Corpus of Spoken Slovak.

Language processing tools

The standard preprocessing steps (tokenization, morphological analysis, POS tagging, lemmatizer) are adequate for Slovak, but the tools could benefit from extended dictionaries.

3.6.6. Slovak resources of greatest interest in the next rounds of selection.

Parallel corpora have been given high priority in the upgrade/extension/alignment process, so more resources of this type are intended to be included in the next batches of resources (to be released in June 2012 and January 2013) and going beyond the end of the project. Another resources we plan to include are collocation dictionaries, valency dictionaries and other existing similar databases.

4. Conclusions

The elaborated methodology is based on the combination of four indicators: general assessment, Total Point Value, Language White Papers and specification the origin of the resources. The priority is given to the general assessment and Language White Papers. The general indicator is a combination of 16 criteria, distributed in three groups. The Total Point Value and Language WhitePapers are based on a numerical assessment of the resources according to qualitative and quantitate criteria. The resource origin is a complementary indicator. Based on the above described methodology the partners assessed the already selected language resources and tools and provided analyses showing both the motivation for particular choices and the identified gaps in the selection.

Table 19. Comparison of the LWPs results for all individual languages

CESAR languages resources	Bulgarian	Croatian	Hungarian	Polish	Serbian	Slovak	Overall average
1. Reference Corpora	4.714	3.286	5.714	3.714	3.429	3.857	4.119
2. Syntax-Corpora (treebanks. dependency banks)	2.143	2.000	4.857	2.857	0.000	2.429	2.381
3. Semantics-Corpora	3.429	0.000	4.143	1.857	0.000	0.000	1.572
4. Discourse-Corpora	1.429	0.000	0.000	1.143	0.000	1.857	0.738
5. Parallel Corpora. Translation Memories	2.429	2.429	5.714	3.857	2.571	2.286	3.214
6. Speech-Corpora (raw speech data. labelled/annotated speech data. speech dialogue data)	2.286	3.000	2.571	1.857	2.857	2.857	2.571
7. Multimedia and multimodal data (text data combined with audio/video)	1.000	2.571	0.571	0.714	1.571	2.143	1.428
8. Language Models	1.571	0.000	4.714	1.286	2.286	2.714	2.095
9. Lexicons. Terminologies	3.571	3.286	4.000	3.286	3.143	3.143	3.404
10. Grammars	2.571	0.000	4.286	2.857	0.714	2.000	2.071
11. Thesauri. WordNets	4.000	2.714	3.429	3.714	3.000	2.857	3.286
12. Ontological Resources for World Knowledge (e.g. upper models. Linked Data)	2.000	0.000	2.429	1.857	0.714	0.000	1.167

The presented table summarises the data for language resources provided in the Language White papers. It shows that relatively equal results for all languages are visible in categories: Reference corpora, Parallel corpora, Speech corpora, Lexicon, Terminoloiges, and Thesauri, WordNets. On the other hand, the table also indicates the gaps of the resources for the target languages: Discourse Corpora, Semantics-Corpora, Multimedia and multimodal data and Ontological Resources for World Knowledge. It is not realistic to expect to fill the gap in the scope of the CESAR project but the clear understanding of necessities and clear definition of the future directions is of great importance.

The report illustrates how the adopted methodology and criteria are applied for each individual language: Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak. It gives an extensive overview and assessment of the selected language resources for the every language and identifies gaps in the provision of the language resources. For each language a profound analysis of the set of already selected resources and tools is performed. The analysis leads to the conclusion what kind of resources should be of the greatest interest in next rounds of selection.

ABBREVIATIONS

Abbreviation	Term/definition
TPV	Total Point Value
PV	Point Value
BLARK	Basic LAnguage Resources Kit
LWP	language White Paper

Table 20. Abbreviations