Detecting Gaps in Language Resources and Tools in the Project CESAR

Marko Tadić[1], Tamás Váradi[2], Radovan Garabík[3], Svetla Koeva[4], Maciej Ogrodniczuk[5], and Duško Vitas[6]

[1] University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb
[2] Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest
[3] Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava
[4] Institute for Bulgarian Language Prof Lyubomir Andreychin, Bulgarian Academy of Sciences, Sofia
[5] Institute of Computer Science of the Polish Academy of Sciences, Warsaw
[6] University of Belgrade, Faculty of Mathematics, Belgrade

Abstract. In this paper the first preliminary results of the analysis of marks collected within the tables of META-NET series of Language White Papers of CESAR project languages are demonstrated. Although they are preliminary results, we can consider them useful for showing us where real gaps in language resources and tools can be detected.

1 Introduction

This paper presents the first preliminary analysis of marks collected within the META-NET series of the Language White Papers (LWP) concerning the languages involved in the CESAR project [1]. The CESAR project is part of the META-NET Network of Excellence and its purpose is to provide the necessary input regarding the language resources and language tools and/or services for languages included in the project, namely, Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak. Instead of producing another "vertical" survey of existing language resources and tools for each language separately, we wanted to turn our viewpoint into a "horizontal" direction that would give us the view on the situation within each category for all CESAR languages, thus pointing us to the area in which the project has to put more effort. The paper is organised as follows: in the section 2 we discuss the data source, in section 3 the results for languages resources are given and discussed, in section 4 we present the results for language tools and discuss them, while the section 5 gives the conclusion.

2 Collecting Data

The first source of data for our analysis are the tables for individual languages produced by the subjective marks given for each of predefined categories. Within the META-NET campaign for producing Language White Papers for 30 European languages in Spring 2011, a collection of

marks given by selected national experts from the LRT field was prepared in the form of tables. One can argue that this procedure is highly dependent on the subjectivity of persons giving the marks, as well as on the availability and reliability of the information for different resources, but the META-NET collecting procedure requested that marks should be given by several experts and then averaged. We can not investigate whether this procedure was respected completely – this was left to the national representatives within the CESAR project and META-NET as a whole to check – so we have taken over the collected marks and did the analysis for the CESAR languages. Also, a list of categories could be speculated upon, but at this moment we have accepted them as they are and we shall see whether this list will be submitted to any reshaping.*
We have taken the marks from the tables of first, unpublished versions of the respective Language White Papers [2,3,4,5,6,7] and processed them in a manner that for each given LRT category we calculated an average of all marks.**
All averages were then mapped to a single space where marks for each category were joined with the language identifier. The same procedure was applied for another type of calculation that included the overall sum of all marks in an individual category instead of their average. As comparison of data produced by these two methods yielded no significant differences between the general shape of results in these two calculations, we selected only one of them – the average of marks. In the rest of the paper all marks regarding individual languages are averaged in the way described above. Having marks spread in this way we could immediately spot the categories in which most of the CESAR languages had very low marks.

3 Results for Language Resources

The results for language resources were produced separately from the language tools/technologies/applications not just because they describe different phenomena or because they have been represented by two different tables in Language White Papers, but also because in this way comparison of results between these two types of LT products can be performed. The results for language resources can be seen at Table 1 and Fig 1. The numbers of categories from the Table 1 are equal to the numbers on the left side of the graphical representation of the table in Fig 1.

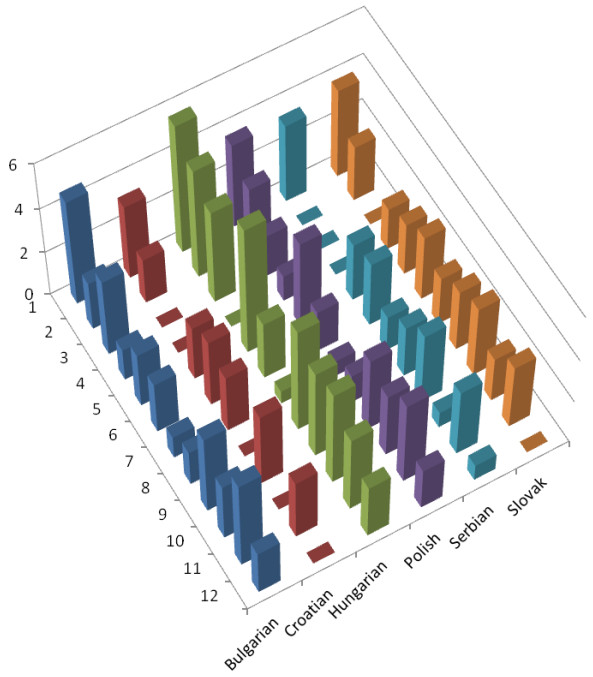
* In the final version of all META-NET Language White Papers, the overall methodology of collecting and merging marks was changed. It was decided that the peer-evaluation of the original fine-grained categories would not be practical and feasible to carry out at the META-NET community level. Therefore the categories were merged and the further process of evaluation and the final decisions at the META-NET meeting in Berlin in 2011 were based on the summary categories.

** Each LRT category was originally marked (on a scale of 0 to 6) for quantity, availability, quality, coverage, maturity, sustainability and adaptability. See the respective tables in Section 4.6 of the individual LWP volumes.

Table 1: Average marks for CESAR language resources

	Bulgarian	Croatian	Hungarian	Polish	Serbian	Slovak	average
1. Reference Corpora	4.71	3.29	5.71	3.71	3.43	3.86	4.12
2. Syntax-Corpora (treebanks, dependency banks)	2.14	2.00	4.86	2.86	0.00	2.43	2.38
3. Semantics-Corpora	3.43	0.00	4.14	1.86	0.00	0.00	1.57
4. Discourse-Corpora	1.43	0.00	0.00	1.14	0.00	1.86	0.74
5. Parallel Corpora, Translation Memories	2.43	2.43	5.71	3.86	2.57	2.29	3.21
6. Speech-Corpora (raw speech data, labelled/annotated speech data speech dialogue data)	2.29	3.00	2.57	1.86	2.86	2.86	2.57
7. Multimedia and multimodal data (text data combined with audio/video)	1.00	2.57	0.57	0.71	1.57	2.14	1.43
8. Language Models	1.57	0.00	4.71	1.29	2.29	2.71	2.10
9. Lexicons, Terminologies	3.57	3.29	4.00	3.29	3.14	3.14	3.40
10. Grammars	2.57	0.00	4.29	2.86	0.71	2.00	2.07
11. Thesauri, WordNets	4.00	2.71	3.43	3.71	3.00	2.86	3.29
12. Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	2.00	0.00	2.43	1.86	0.71	0.00	1.17

Fig. 1: Graphical representation of Table 1



From the Table 1 and also from Fig 1 it is clearly observable which category of LR are deficient. The lowest overall average (0.74) is in

Category 4 Discourse Corpora, but also below average mark 2.00 are Category 3 Semantics-Corpora (1.57), Category 7 Multimedia and multimodal data (1.43) and Category 12 Ontological Resources for World Knowledge (1.17).
What is worth noting is the fact that in half of the categories at least one language has mark 0.00 and there are two categories where three languages have mark 0.00: Category 3 Semantics-Corpora and Category 4 Discourse-Corpora.
Also, a considerable discrepancy between individual languages can be noticed in certain categories, e.g. in Category 3 Semantics-Corpora Bulgarian, Hungarian and Polish have 3.43, 4.14, and 1.86 respectively while Croatian, Serbian and Slovak have 0.00.
If we look at the contents of these categories then some very low marks (e.g. Categories 3 and 4) are explainable by the status of under-resourced languages as more languages exhibit 0.00 there. The opposite case, when only one language had mark 0.00 (e.g. Serbian in Category 2 Syntax Corpora, or Croatian in Category 8 Language Models), can be interpreted as significant deficiency in this type of resource for this particular language. The reasons for this deficiency could be different, starting from researchers' preferences in research priorities, up to insufficient national funding for these resources. However, it is a very good indicator that this type of resources should be developed in the near future for a particular language. Such figures could be helpful in argumentation for future funding applications.
Consistent results over all languages are visible in Categories 1 Reference corpora, 5 Parallel corpora, 6 Speech corpora, 9 Lexicon, Terminologies, and 11 Thesauri, WordNets. This leads to the conclusion that for these types of resources there are good representatives in respective languages and that they reached certain level of maturity. One could argue that this result is to be expected since these are basic language resources and usually development of LT for a certain language starts with them. Also, in some languages the LR&T community goes back to several decades and in spite of usually poor funding from industry, they managed to build basic resources funded from other directions.

4 Results for Language Tools

The results for language tools were produced separately from the language resources following the same procedure of averaging. The results are given in Table 2 and Fig 2. The numbers of categories from the Table 2 are equal to the numbers on the left side of the graphical representation of the table in Fig 2.
The top view over the Table 2 and Fig. 2 can lead us to the general observation that the number of lower grades is higher in the case of language tools compared to language resources for CESAR languages. It is particularly noticeable by the number of marks 0.00, where there are 17 cells (21.79%) with that mark for language tools, while in language resources there were only 11 cells (15.28%).
For particular categories the lowest overall average (0.1) is in Category 5 Advanced Discourse Processing, but also below average mark 1.00

Table 2: Average marks for CESAR language technology (Tools, Technologies, Applications)

	Bulgarian	Croatian	Hungarian	Polish	Serbian	Slovak	average
1. Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	4.00	3.57	4.00	4.57	4.29	3.00	3.90
2. Parsing (shallow or deep syntactic analysis)	3.00	1.57	3.57	3.57	2.43	0.00	2.36
3. Sentence Semantics (WSD, argument structur, semantic roles)	2.43	1.14	1.57	2.14	0.00	0.00	1.21
4. Text Semantics (coreference resolution, context, pragmatics inference)	1.43	0.00	1.29	1.00	0.00	0.00	0.62
5. Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	0.00	0.00	0.00	0.57	0.00	0.00	0.10
6. Information Retrieval (text indexing, multimedia IR, crosslingual IR)	2.00	2.29	0.86	3.29	2.43	2.29	2.19
7. Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	2.29	2.43	5.57	2.57	2.14	1.71	2.79
8. Language Generation (sentence generation, report generation, text generation)	1.43	1.29	0.00	1.14	0.00	0.00	0.64
9. Summarization, Question Answering, advanced Information Access Technologies	1.86	0.29	0.00	1.29	0.71	1.71	0.98
10. Machine Translation	2.29	0.71	4.86	3.29	0.71	1.86	2.29
11. Speech Recognition	2.00	2.57	2.71	2.71	1.14	2.29	2.24
12. Speech Synthesis	2.00	3.57	3.71	4.14	3.29	3.00	3.29
13. Dialogue Management (dialogue capabilities and user modelling)	0.00	1.29	0.00	1.00	0.00	0.00	0.38

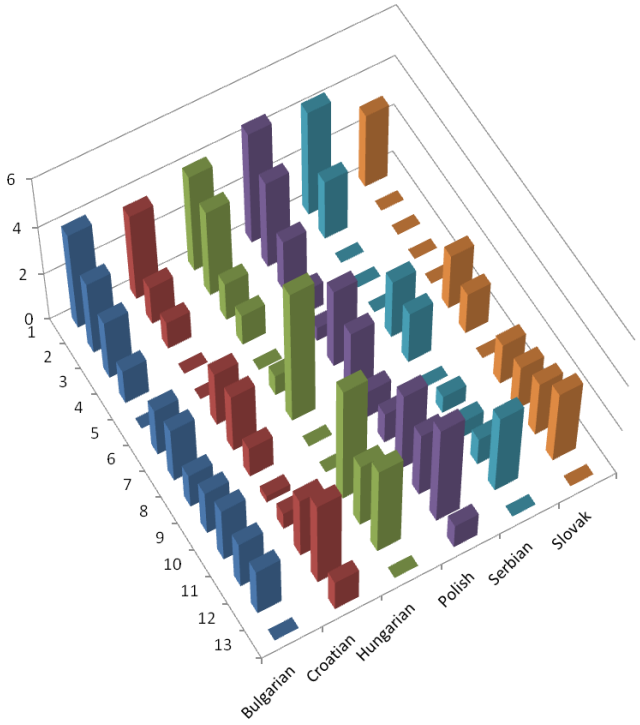
are Category 4 Text Semantics (0.62), Category 8 Language Generation (0.64), Category 9 Summarization. Question Answering. advanced Information Access Technologies (0.98) and Category 13 Dialogue Management (0.38). These numbers tell us that 38.46% of all categories have mark below 1.00 on the scale from 0 to 6 and this is very low.

Also in seven categories (53.85%) at least one language has mark 0.00 and there are categories where four or five languages have mark 0.000.

A considerable discrepancy between individual languages can be noticed only in the Category 2 Parsing where Slovak has 0.00, while all other languages have above 1.50, with the average of 2.36 for the whole category. In other cases there are marks zero for more than one language or the overall average mark is below 1.00. This means that more languages have low marks for many language tools and this clearly defines the under-resourced status of CESAR languages regarding the necessary language tools.

Consistent results over all languages are visible only in Categories 1 Tokenization. Morphology, 7 Information Extraction, and 12 Speech synthesis. Knowing that most of the languages in the CESAR project do have rather complex inflectional and derivational morphology (e.g. noun

Fig. 2: Graphical representation of Table 2



inflection complexity starts from Bulgarian where there are no cases, just singular and plural wordforms, to other Slavic CESAR languages, having, usually seven cases in singular and plural, up to the extremely complex Hungarian with about twenty cases in both numbers), it is no surprise that the majority of efforts of development of LT were concentrated previously in Category 1. Also, Category 7 Information Extraction is the next expected field where the fundamental findings from Category 1 can find their application, particularly with the NERC systems that could more easily find their market niche than other types of tools. Speech synthesis is also expected in this bunch since it is easier to start with synthesis than speech analysis and thus it is the usual direction of development in speech processing for a given language.

Like in the case of language resources, the detected gaps are very good indicators that this type of tools/services should be developed in the near future for a particular language. The above findings could be used as very strong arguments in requests for additional funding at the national level.

5 Conclusion and Future Directions

We have just presented the first preliminary results of the analysis of marks collected within the tables of META-NET series of Language White Papers for the languages of the CESAR project. Although they are preliminary results, we can consider them useful for showing us where real gaps in language resources and tools can be detected. Since we were aware that the CESAR languages are under-resourced compared to e.g. English or German, we were prepared for some low grades, but some categories had marks below any expectation.

The standard preprocessing steps (tokenization, morphology, shallow parsing etc.) are more-or-less completed, but the more difficult semantics and discourse analysis need further research. The higher the linguistic processing level the lower the scores are, as can be seen in the first five rows of Table 2 (Tokenization. Morphology: 3.91, Parsing: 2.36, Sentence Semantics: 1.21, Text Semantics: 0.62, Advanced Discourse Processing: 0.10). This is justified by the fact that syntax and semantics are more difficult to process than morphology. The more semantics a tool takes into account, the more difficult it is to find the right data and more efforts for supporting deep processing are needed. Semantic tools and resources are scored very low. Thus, programs and initiatives are needed to substantially boost this area both with regard to basic research and the development of annotated corpora.

One of the future directions could involve studying the discrepancy between the existing tool and the non-existing resource for a combined set of categories that depend on each other, e.g., in language resources Category 2 Syntax corpora and in language tools Category 2 Parsing.

Since this first analysis was not done using any elaborated statistical instruments, but simply by comparison of averages of marks, it might happen that the results obtained by a proper statistical treatment (median, standard deviation, hypothesis testing, etc.) will be somewhat different, at least the possible bias in giving marks for certain categories and/or languages could be avoided.

Also, a set of categories can be statistically verified for their significance and this may lead to joining or disjoining of some categories making the grid for marking more dense or coarse.

It should be noted that, for reasons mentioned in Section 2, the categories in the tables presented here and the categories in the tables in the final version of Language White Papers are not mutually comparable. However, we have shown that the more dense grid presented here allowed the precise detection of the weak spots in the development of LRT for each of CESAR languages.

References

1. Váradi, T.: Veni, Vidi, Vici: The Language Technology Infrastructure Landscape after CESAR. In Gajdošová, K., Žáková, A., eds.: Natural Language Processing, Corpus Linguistics, E-Learning, Bratislava, RAM-Verlag (2013) 261 – 278

2. Blagoeva, D., Koeva, S., Murdarov, V.: Българският език в дигиталната епоха – The Bulgarian Language in the Digital Age. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer (2012) Available online at http://www.meta-net.eu/whitepapers.
3. Tadić, M., Brozović-Rončević, D., Kapetanović, A.: Hrvatski Jezik u Digitalnom Dobu – The Croatian Language in the Digital Age. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer (2012) Available online at http://www.meta-net.eu/whitepapers.
4. Simon, E., Lendvai, P., Németh, G., Olaszy, G., Vicsi, K.: A magyar nyelv a digitális korban – The Hungarian Language in the Digital Age. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer (2012) Available online at http://www.meta-net.eu/whitepapers.
5. Miłkowski, M.: Język polski w erze cyfrowej – The Polish Language in the Digital Age. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer (2012) Available online at http://www.meta-net.eu/whitepapers.
6. Vitas, D., Popović, L., Krstev, C., Obradović, I., Pavlović-Lažetić, G., Stanojević, M.: Српски језик у дигиталном добу – The Serbian Language in the Digital Age. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer (2012) Available online at http://www.meta-net.eu/whitepapers.
7. Šimková, M., Garabík, R., Gajdošová, K., Laclavík, M., Ondrejovič, S., Juhár, J., Genči, J., Furdík, K., Ivoríková, H., Ivanecký, J.: Slovenský jazyk v digitálnom veku – The Slovak Language in the Digital Age. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer (2012) Available online at http://www.meta-net.eu/whitepapers.