Making Monolingual Corpora Comparable: a Case Study of Bulgarian and Croatian

Božo Bekavac*, Petya Osenova†, Kiril Simov†, Marko Tadić*

*Institute of Linguistics, Faculty of Philosophy,
University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
bbekavac@ffzg.hr, marko.tadic@ffzg.hr

†BulTreeBank Project
LML, Bulgarian Academy of Sciences
Acad. G. Bonchev St. 25A, 1113 Sofia, Bulgaria
petya@bultreebank.org, kivs@bultreebank.org

Abstract

This paper describes the first steps towards the creation of a Bulgarian-Croatian comparable corpus. Its base are two newspaper sub-corpora from larger reference corpora of Bulgarian and Croatian. In the beginning we rely on more extralinguistically-oriented, but methodologically cleaner parameters of similarity like: specific topics, pre-defined time span and data size. The idea of 'light' and 'hard' comparable corpora is introduced. At this stage we aim at producing a 'light' bilingual comparable corpus. The algorithm for identifying lexical similarity and aligning linguistic units is presented, and the initial experiments are outlined.

1. Introduction

The idea for comparable corpora is still somewhat new and therefore, still underexplored. There are only few examples of such corpora already collected and available: e.g. the newspaper Portuguese corpus (CETEMPublico) and Reuter's English corpus; ICAME corpora, among others (for more details see (Maia 2003))

Some attempts are made for automatic alignment of monolingual comparable corpora (Barzilay and Elhadad 2003). In this respect two techniques are applied: *vertical paragraph clustering* and *horizontal paragraph mapping*. The former is needed for clustering the paragraphs in each corpus by the same type of information. In this method the specific names, dates and numbers are ignored. The latter is used for mapping pairs of paragraphs from both corpora.

According to the EAGLES (EAGLES 1996) specifications the idea behind comparable corpora is: "to compare different languages or varieties in similar circumstances of communication, but avoiding the inevitable distortion introduced by the translations of a parallel corpus." Hence, the advantages of producing such corpora are as follows:

1. they do not depend on the quality or the specific cultural nuance of translations. On the contrary, they view texts as original samples in the source language culture.

2. usually original texts on a particular subject are more easily available than good translations.

3. versatility, i.e. they can be used for wider range of NLP tasks than parallel corpora.

Without underestimating the importance of parallel corpora, we think that more effort should be invested in the idea of comparable corpora.

In this paper, we describe a starting initiative for the creation of a Bulgarian-Croatian comparable corpus, given two newspaper sub-corpora from Bulgarian and Croatian general reference corpora. The objective of most comparable corpora is to concentrate on producing small corpora in specific areas. However, our aim at the moment is to map the two sub-corpora according to a pre-defined set of common criteria.

We have decided to start with more extra-linguistically-oriented, but thus methodologically cleaner and safer, parameters of similarity. Here are the parameters:

1. from content point of view: **domain**: daily newspapers with great social impact, we concentrate on one newspaper; **topic**: external politics and sport as most 'eurospeak' and culture-independent areas;

2. from formal point of view: **size**: 1 million tokens; **format**: XML, TEI 1st level structure markup (up to the paragraphs); **time span**: 2001.

The structure of the paper is as follows: Section 2 and 3 describe briefly the Bulgarian reference corpus and its newspaper subpart, and the Croatian reference corpus and its subpart, respectively. In section 4 the idea of 'light' comparability is introduced, the starting algorithm is outlined and the initial experimental work is presented. The last section summarizes our conclusions, describes the directions for future work and focuses on the advantages of the comparable Bulgarian-Croatian corpus.

2. The Bulgarian Corpus

2.1. The General Corpus

It is intended to yield the size of a national corpus, that is, 100 million running words. Since the data are gradually annotated, its status at the moment is approximately as follows:

Nearly 90 million running words are collected from different sources in HTML and RTF formats. In order to compile a representative and balanced corpus of Bulgarian texts, we tried to gather a variety of different genres: 15 % fiction, 78 % newspapers and 7 % legal texts, government bulletins and others.

About 72 million running words are converted into XML documents, marked up in conformance with the TEI guidelines. This conversion is automatic: for each source of text we developed a separate tool for extraction of the relevant information like the text itself, but also the author

information, genre classification (where it is available), and other meta-information. The tools are implemented in Prolog and the CLaRK system (Simov et. al. 2001).

2.2. The Subcorpus

The Bulgarian part of the comparable corpus consists of articles from the daily newspaper "Sega" from 2001. It has altogether 393757 tokens in the articles from foreign policy rubric and the articles from sports rubric. The texts are part of the BulTreeBank corpus. The originally HTML source documents were converted into XML format and validated in the CLaRK system.

3. The Croatian Corpus

3.1. The General Corpus

The Croatian National Corpus (HNK) is planned to achieve the size of 100 million tokens by 2004. In the current stage the texts collected encompass more than 130 million of tokens but they are not in full accordance with predefined structure of genres and text-types: 74% faction (newspapers, magazines, textbooks, law, etc), 23% fiction (novels, stories, essays, etc) and 3% mixed texts (memoirs, chronicles, etc). For more details see (Tadić 2002).

Texts in the overall size of 20 million tokens are converted to XML documents marked up in conformance to TEI level 1 recommendations. The proprietary tools for automatic conversion 2XML has been developed which allows batch conversion in the two-step process with use of user-defined scripts. The test version (ca 12 million tokens) is freely searchable at the `http://www.hnk.ffzg.hr`.

3.2. The Subcorpus

The Croatian part of the comparable corpus consists of articles from the daily newspaper "Vjesnik" from 2001. It comprises 3427 articles from foreign policy rubric (1.5 million tokens), and 3376 articles from sports rubric (1.6 million tokens). The texts are collected within the scope of the larger corpus project – Croatian National Corpus. The source documents in HTML format were automatically converted into XML with our own custom made tool 2XML (see more in (Tadić 2002)). The articles are marked up in the conformance with the TEI guidelines and prepared for the importing into CLaRK where they were processed further.

4. Towards Comparability of the Corpora

We would like to introduce the notion of two levels of comparability of corpora. Let's call them "light" and "hard" comparability and therefore corpora can be comparable in "light" or "hard" way.

The first type of comparability is characterised by having corpora from two (or more) languages composed according to the same principles (i.e. corpora parameters) which are defined by extralinguistic and extratextual features such as size, time-span, text genres (newspaper rubrics), gender and/or age of the authors etc. Nowadays this kind of comparable corpora can be acquired quite easily from already existing corpora or from other e-text sources.

The second type of comparability is dependent on already collected and established "lightly" comparable corpora. It is derived from them by applying certain language technology tools/techniques and defined parameters of their usage to find out which documents in lightly comparable corpora really deal with the same or similar topic. Subset of lightly comparable corpora which has been selected by those tools/techniques can be regarded as a "hard" comparable corpora. The possible techniques could be simple comparison of frequency lists of lemmas and/or collocations; named entity recognition, classification and comparison; document classification; term extraction comparison etc.

As it was mentioned in the introduction, we decided to start with samples from the newspaper monolingual subcorpora of the two languages. First, we compared samples of the newspapers *Sega* (Bulgarian (bg)) and *Vjesnik* (Croatian (hr)) from year 2001. Our work was facilitated by the following facts: the newspaper data is thematically structured, the percentage of international lexica is very high, the same events are discussed, the mark-up is XML. The considered sections were restricted to the foreign policy pages: 'Chuzhbina' (bg), 'Vanjska politika' (hr) and Sport (bg and hr).

4.1. Starting Algorithm

We rely on two types of prerequisites: *extralinguistic* (popular newspapers; identical dates, which presuppose identical events) and *linguistic* (the same pre-defined rubrics, which presuppose the similar topical structure and ensure a high lexical similarity). In this respect our approach differs from other approaches, which aim at aligning sentences with little surface resemblance (Barzilay and Elhadad 2003).

Our starting algorithm followed the steps below:

1. Manual mapping of article pairs according to their headings and/or key words. It turned out that only the information in the headings is not reliable enough for handling the article alignment properly. The reason is that the headings do not always focus on the same part of the event. In such cases, the lexical similarity within the articles is measured. Hence, the next step was to find some other supporting techniques and automate the mapping procedure.

2. First, we relied on Bulgarian-English and Croatian-English lexicons for discovering lexical equivalents in the headings and in the texts. As we expected event-specific information, it depended heavily on named-entities. Our reasons for choosing English as a mediating language ('lingua franca') are the following:

   - The rubrics that have been selected for comparison are internationally oriented and a vast amount of the texts have been translated from English;
   - There are very well elaborated bilingual resources in both languages with respect to the English language. It is better to re-use them than to create a new large database;

- The corpus can be easily extended to cover English and other languages through the mediation of English.

Needless to say, the above step cannot substitute the need of bi-directional Bulgarian-Croatian and Croatian-Bulgarian dictionaries. It is crucial not only for mapping the common words, but also: for mapping Bulgarian names in Croatian texts and vice versa; when handling nationality-specific realia and for word-sense disambiguation. Thus, we need several kinds of dictionaries as: morphological one, named-entities one and explanatory one.

This step was connected with the creation of two types of lexicons: (1) a common lexicon and (2) a named-entities lexicon. All the words or phrases in the lexicons were weighed according to the following criterion: all the unique objects were assigned 1 (for example, the names of countries or politicians), all the descriptions ('the president of the USA') were assigned weights 0.8, all stop words were assigned weight 0. Additionally, we relied on encyclopedic knowledge concerning the political and sports domains. In this respect we could predict what named-entities to be expected in the texts. Consequently, we avoid vertical clustering and apply only horizontal mapping, i.e. two articles are matched if their headings/texts show lexical similarity.

3. If the lexical similarity within the texts is high, then the alignment could be further refined to the sub-paragraph level. One indicator of such a high similarity might be the common source (for example, Reuters) or the common target location (for example, Germany).

   Note that for a more refined alignment we should employ some paraphrasing techniques. This is needed because certain specific events are described more briefly in one newspaper in comparison with the other. For example, the aligned texts about Biljana Plavshich's arrest are of different sizes despite the fact that they have the same information source.

All the steps, mentioned above, had to be tested against their applicability. For that reason we have performed some statistics over the newspaper rubrics of 10 days from January, 2001. The results and comments are presented in next subsection.

4.2. The Statistics

We aimed at deriving three types of information: (1) token frequency, (2) type frequency and (3) distribution. It is worth noting that the tokens in Croatian texts are nearly twice as the Bulgarian ones (21 034 tokens vs. 12 661 tokens). First, some observations were done over the first 1000 tokens. Then, the tokens below this number were also considered.

Excluding the stop words, we have observed the following:

1. *Token frequency.* Within the first most frequent 1000 tokens the lexis, which characterizes the two domain-specific field, shows higher similarity. Note that some of the matches presented here are between word forms, not between lemmas. For example, named entities and words connected to politics: 'Clinton' (English (eng)) - 'Klintyn' (bg)[1] (26 occurrences) vs. 'Clinton' (hr) (28 occurrences); 'Moscow' (eng) - 'Moskva' (bg) (25 occurrences) vs. 'Moskvi' (hr) (23 occurrences) and 'Moskve' (hr) (4 occurences); 'Europe' (eng) - 'Evropa' (bg) (17 occurrences) vs. 'Europi' (hr) (18 occurrences), 'minister' (eng) - 'ministyr' (bg) (24 occurrences) vs. 'ministar' (hr) (25 occurrences). Typical for sport: 'league' (eng) - 'liga' (bg) (32 occurrences) vs. 'liga' (hr) (24 occurrences), 'coach' (eng) - 'trenyor' (bg) (33 occurrences) vs. 'trener' (hr) (38 occurrences) etc.

   Other tokens which show a high frequency similarity fall into the following groups: (1) verbs of saying: 'said' (eng) - 'zayavi' (bg) (73 occurrences) vs. 'rekao' (hr) (102 occurrences); (2) modal verbs: 'can' (eng) - 'mozhe' (bg) (57 occurrences) vs. 'može' (hr) (86 occurrences); 'must' (eng) - 'tryabva' (bg) (48 occurrences) vs. 'trebao' (hr) (53 occurrences); (3) relatives: 'where' (eng) - 'kydeto' (bg) (69 occurrences) vs. 'gdje' (hr) (79 occurrences); temporal and quantity measurements: 'year' (eng) - 'godina' (bg) (75 occurrences) vs. 'godina' (hr) (66 occurrences), 'dollars' (eng) - 'dolara' (bg) (39 occurrences) and 'dolari' (bg) (3 occurrences) vs. 'dolara' (hr) (33 occurrences).

2. *Type frequency.* We have not performed automatic type frequency, because lemmatization is needed first. As type frequency depends on adding morphological knowledge, it is left for the next stage, in which the comparable texts will be linguistically processed.

3. *Distribution.* The distribution of the tokens and types can be divided into two kinds. The first one refers to the division of the elements into *comparable* (international lexis and key words) and *non-comparable* (national realia names).

   One interesting observation within the comparable units is that the source 'Reuters' is explicitly stated 110 times in Bulgarian texts, while in Croatian texts it is mentioned only 13 times. Thus it turns out that the explicit presence of the source is not a reliable indicator for lexical similarity.

   Within the domain of the non-comparable units the name Bulgaria (Bylgariya) was mentioned 29 times in the Bulgarian texts. In parallel, Croatia (Hrvatska) was mentioned 30 times in Croatian texts.

   The second division takes into account the different distribution of the tokens from the same type and the distribution of different types. Thus, for example, in Croatian texts Moscow has 23 occurrences in the form 'Moskvi' and only 4 occurrences in the form 'Moskve'

[1] All Bulgarian examples are transliterated within the Latin alphabet.

or in Bulgarian texts the lemma 'dollar' has 39 occurrences in its count form 'dolara' and only 3 occurrences in its plural form 'dolari' (see above). This fact should not be ignored, because the generalization over certain token preferences can give clues for the structure of the media language in the political and sports domains. As a result, other control techniques for alignment can be introduced. Concerning the type distribution, less frequent are named-entities and common words that are not so specific for the considered domains. For example, all the word forms of the word 'center' show a low frequency distribution - in Croatian each token has frequency 3, while in Bulgarian there are some ignorable differences - the lemma has occurred 6 times, the form with short definite article - 7 times, the form with the full definite article - 2 times, plural form - just once.

4.3. Experiment Description

At the start we have compared and manually aligned newspaper issues of two days (10 and 11 January 2001) - 19 articles for Croatian and 27 articles for Bulgarian within the CLaRK system. Four matches were detected, which is around 20 % from the available data. Two of the headings were matchable ('Grymna elektrocentrala v germanski koncern' (bg) vs. 'Niz eksplozija u njemackoj elektrani' (hr); 'Bilyana Plawshich se predava sama v tribunala na OON za woenni prestypleniya' (bg) vs. 'Bivša predsjednica Republike Srpske odlućila se dragovoljno predati sudu u den Haagu' (hr)) and two were not directly matchable ('Ludata krawa "posturi" germanskoto pravitelstvo' (bg) vs. 'Ministrice zdravstva i poljoprivrede obećavaju povratak povjerenja potrošaća u mesnu industriju' (hr) etc.). It confirms the fact that the techniques for alignment have to take into account not only the relations between structurally identical texts, but also the relations between structurally different pieces of texts, such as headings and normal text. Thus, we first rely on larger units of text to be aligned (whole articles) before matching paragraphs into paragraphs and sentences into sentences. This strategy is justified by the observation that in contrast to the parallel corpora, in comparable ones the information flow seems to be non-homogenously distributed in headings and texts. So, the relations between the elements are not viewed as 'onto' relation, but rather 'into' relation, i.e. as a net of relations from everywhere to everywhere - from Bulgarian heading to Bulgarian body text, from Bulgarian heading to Croatian body text, from Croatian body text to Bulgarian body text etc.

Next, the newspaper issues for 10 days of January 2001 were selected for token frequency statistics. The articles were unified, tokenized and, after the application of the statistical module, they were sorted by tokens. Even though the number of the tokens in the Croatian newspaper outcomes the number of tokens in the Bulgarian one, the results were promising with respect to the frequency token matches. More work is to be done at the level of types.

5. Conclusions and Outlook

We described the first steps towards the creation of a Bulgarian-Croatian comparable corpus from existing reference corpora. Since two languages are genetically and geographically close, the relatively similar degree of internationalization and coverage of the events important for both countries is ensured. We started with samples from the newspaper sub-corpora, because they seem to be easily comparable. Two conclusions can be drawn at this stage:

1. knowledge-based resources are required for more precise mappings, and
2. the notion of comparability in contrast to parallel corpora presupposes many-to-many relations between the units and thus, becomes a real challenge for the corpora developers.

We envisage to continue our joint work in the following directions:

1. to extend the experiments over other domains and on more data,
2. to automate the alignment procedure starting from general mappings and aiming at more precise ones,
3. connecting the event structure of the articles with the temporal frame, i.e. mapping temporal expressions with respect to the date and the year of the issue. In this way, for example, descriptions like 'the president of the USA' will be anchored to the right name.

Once created, the corpus will allow the linguistic judgments with corpus data in more controlled conditions. It will also be useful for testing of the same language tools on different languages, information and term extraction, automatic lexicon building, etc. It might be useful not only to linguists but also to social anthropologists, sociologists of culture, Central and Eastern European studies of any kind etc.

6. References

Barzilay R. and Elhadad N. 2003. *Sentence Alignment for Monolingual Comparable Corpora.* In: *Proc. of EMNLP.*

Maia. 2003. *What are Comparable Corpora? Multilingual Corpora: Linguistic Requirements and Technical Perspectives.* In: *Proc. of A pre-conference workshop, Corpus Linguistics 2003 Conference.* England.

EAGLES. 1996. *Expert Advisory Group on Language Engineering Standards Guidelines.*

Kiril Simov, Zdravko Peev, Milen Kouylekov, Alexander Simov, Marin Dimitrov, Atanas Kiryakov. 2001. *CLaRK - an XML-based System for Corpora Development.* In: *Proc. of the Corpus Linguistics 2001 Conference.* England. pp 558–560.

Marko Tadić. 2002. *Building the Croatian National Corpus.* In: *Proc. of the LREC conference.* Canary Islands, Spain. pp 441–446.

Wolfgang Teubert. 1995. *Language Resources: The Foundations of a Pan-European Information Society.* In: *Proc. of the 1st TELRI seminar.* IDS, Mannheim. Germany.