Extracting Translation Lexicons from Bilingual Corpora: Application to South-Slavonic Languages

Preslav Nakov
University of California at Berkeley
EECS dept., CS division
Berkeley, CA 94720
nakov@cs.berkeley.edu

Veno Pacovski
Sts. Cyril & Methodius University
Dept. of Mathematics and Informatics
Skopje, Macedonia
pacovski@gf.ukim.edu.mk

Elena Paskaleva
Bulgarian Academy of Sciences
CLPP, Linguistic Modeling dept.
25A Acad. G. Bonchev Str.
Sofia, Bulgaria
hellen@lml.bas.bg

Abstract

The paper presents a novel approach for automatic translation lexicon extraction from a parallel sentence-aligned corpus. This is a five-step process, which includes cognate extraction, word alignment, phrase extraction, statistical phrase filtering, and linguistic phrase filtering. Unlike other approaches whose objective is to extract word or phrase pairs to be used in machine translation, we try to induce meaningful linguistic units (pairs of words or phrases) that could potentially be included as entries in a bilingual dictionary. Structural and content analysis of the extracted phrases of length up to seven words shows that over 90% of them are correctly translated, which suggests that this is a very promising approach.

Keywords

Lexicons, parallel corpora, machine translation, word alignments, lexicography, cognates, competitive linking, longest common subsequence ratio.

1 Introduction

In the present paper, we describe a novel approach for automatic translation lexicon extraction from a parallel sentence-aligned corpus, trying to induce meaningful linguistic units – pairs of words or phrases – that could potentially be included as entries in a bilingual dictionary.

The method is relatively language-pair independent[1]; it does not require sophisticated linguistic analysis, which makes it particularly suitable for languages with scarce resources, for which using large corpora reflecting contemporary language usage is the ultimate way to go.

Here we apply it to a pair of closely-related South-Slavonic languages – Bulgarian and Macedonian – which is of particular interest for a variety of reasons: historical, political, and linguistic[2].

The remainder of the paper is organised as follows: Section 2 introduces the method giving a detailed description of its five steps, section 3 describes the experiments and presents the results of the evaluation, section 4 points to some related work, and section 5 concludes with possible directions for future work.

2 Method

In our approach, the process of construction of a bilingual lexicon consists of the following five steps:

1. Cognate extraction.
2. Word alignments.
3. Phrases extraction.
4. Statistical phrase filtering.
5. Linguistic phrase filtering.

Each step is explained in detail below.

2.1 Step 1: Cognate Extraction

Since our training corpus is relatively small, we extracted and used potential cognates in order to bias the training of the IBM word alignment models (see below).

Traditional linguistics defines cognates as words derived from a common root [4]. Following previous

[1] Given a pair of languages, the method only requires a parallel sentence-aligned corpus for that pair as well as language-specific lists of stopwords for the two languages.

[2] There is a heated linguistic (and political) debate about whether Macedonian represents a separate language or is a regional literary form of Bulgarian. Since no clear criteria for distinguishing a dialect from a language exist, linguists remain divided on that issue.

researchers in computational linguistics [3, 14, 16], we adopt a simplified definition, which ignores origin, defining cognates as words in different languages that are translations and have a similar orthography.

Following Melamed'95 [15], we measure the orthographic similarity using *longest common subsequence ratio* (LCSR), which is defined as follows:

$$\textsc{lcsr}(s_1, s_2) = \frac{|\text{LCS}(s_1,s_2)|}{\max(|s_1|,|s_2|)}$$

where $\textsc{lcs}(s_1, s_2)$ refers to the *longest common subsequence* of s_1 and s_2, and $|s|$ is the length of s.

Consider for example the Bulgarian *b. казарма* and its Macedonian translation *m. касарна* (mil. '*barracks*'). We have:

$$\textsc{lcs}(\textit{казарма}, \textit{касарна}) = \textit{каара}$$

and therefore:

$$\textsc{lcsr}(\textit{казарма}, \textit{касарна}) = 5/7$$

Following Nakov & al.'07 [18], we use the LCSR similarity measures in combination with *competitive linking* [17] in order to extract potential cognates from our parallel sentence-aligned Bulgarian-Macedonian corpus. Competitive linking assumes that given a source Bulgarian sentence and its Macedonian translation, each source word is either translated with a single target word or is not translated at all. Given a sentence pair, the similarity is calculated for all Bulgarian-Macedonian word pairs[3], which induces a fully-connected weighted bipartite graph. Then a greedy approximation to the maximum weighted bipartite matching in that graph is extracted as follows: First, the most similar pair of unaligned words is aligned and both words are discarded from further consideration. Then the next most similar pair of unaligned words is aligned and the two words are discarded, and so forth. The process is repeated until there are no unaligned words left or until the maximal word pair similarity falls below a pre-specified threshold θ ($0 \leq \theta \leq 1$), which could leave some words unaligned[4]. As a result we end up with a list C of potential cognate pairs. Table 1 shows sample cognates extracted using the described method.

Bulgarian	**Macedonian**	**English**	LCSR
конфликт	конфликт	conflict	1.0
книги	книги	books	1.0
речник	речник	dictionary	1.0
стара	стара	old	1.0
автор	автор	author	1.0
тероризъм	тероризам	terrorism	0.89
държава	држава	state	0.86
такава	таква	such (fem.)	0.83
военно	воено	military	0.83
исторически	историски	historical	0.82
Балканите	Балканот	the Balkans	0.78
отхвърлят	отфрлат	they reject	0.78
свят	свет	world	0.75
очакваха	очекуваа	they expected	0.75
показват	покажуваат	they show	0.70
мисля	мислам	I think	0.67
искат	сакаат	they want	0.67
разрешаване	решавање	solution	0.64

Table 1: Sample cognates. *Extracted on step 1 using* LCSR *and competitive linking.*

2.2 Step 2: Word Alignments

Following [2, 13, 18] we filter out the duplicates from the list of cognates C, and we add the remaining cognate pairs twice as additional "sentence" pairs to the bilingual sentence-aligned corps in order to bias the subsequent training of the IBM word alignment models. This technique has been shown to improve the alignment quality for another pair of closely related Slavonic languages: Bulgarian-Russian [18].

We use IBM model 4 [5] to generate two word-level alignments for this augmented corpus. The IBM word alignment models assume that each source word is connected to exactly one target word; the case of source words with no translation is handled by assuming an initial *null* word in the target sentence to which they are connected. Therefore, the IBM models are directed M:1 models. In order to generate undirected M:M word-level alignments, we first generate two directed alignments, Bulgarian→Macedonian and Macedonian→Bulgarian, and we subsequently symmetrize them using the *interect+grow* heuristic described in [20], which starts with the intersection of the two alignments and then explores the space between the intersection and the union. The resulting M:M word-level alignments are subsequently used for phrases extraction, as described below.

In fact, the word alignments themselves can and have been widely used for automatic word-level translation lexicon extraction. For example, the word links in the intersection of the two directed alignments are typically over 95% correct. Therefore, using the word-level lexical probabilities and suitable thresholds, very accurate word-level translation lexicons can be extracted. However, it is not that easy to extract high-quality phrase-level translation lexicons with a dictionary-like quality.

2.3 Step 3: Phrases Extraction

The above-descried word-level alignments are used to extract phrase-level translations pairs with the *alignment template approach* of Koehn&al.'03 [12]. The approach extracts pairs of contiguous pieces of text, called *phrases*, which do not necessarily represent linguistic units. Given a sentence pair, the phrases are required to be consistent with the word-level alignments in that sentence pair in the sense that the words from the source phrase should only be word-aligned to words from the target phrase, and vice versa, the words from the target phrase should be word-aligned to words from the source phrase only. In addition, unlike the IBM word alignment models, empty phrases

[3] Due to their special distribution, stopwords and short words (one or two letters) are not used in competitive linking.

[4] In our experiments, we use the value $\theta = 0.58$, which has been suggested by Kondrak&al.'03 [13] after multiple experiments with ten European languages from the Europarl corpus [10].

are not allowed. All phrases meeting these constraints from all sentences are extracted, with an additional limitation on the maximum phrase length[5]. The extracted phrase pairs are then scored and the following conditional probabilities are calculated for each Bulgarian-Macedonian pair (b, m):

- forward phrase translation probability: $\Pr(m|b)$;
- reverse phrase translation probability: $\Pr(b|m)$;

In the process of evaluation, we use the minimum of these conditional probabilities as a measure of the phrase pair quality.

2.4 Step 4: Statistical Phrase Filtering

While the above-described translation table produced with the alignment template approach is the backbone of the state-of-the-art phrase-based statistical machine translation model [9], it contains a lot of noise. This is especially true for the low probability phrase pairs. However, it is not easy to filter the phrase table; the most obvious approaches negatively affect the machine translation quality, which is an indication that many useful phrase pairs are lost in the filtering process.

A notable exception is a recent method described by Johnson&al.'07 [8]. Using *Fisher's exact test* [1] and a natural threshold (which excludes all pairs of phrases such that both the source and the target phrases occur exactly once in the parallel corpus), they are able to reduce the size of the phrase table by about 90% without adversely affecting the translation quality as measured by BLEU score [21] in some cases even a small improvement in BLEU is reported.

Indeed, a quick scan through the filtered phrase table reveals that it looks much better than the original one. In our experiments, we have found that this filtered table is a better source for automatic lexicon extraction than the original table.

2.5 Step 5: Linguistic Phrase Filtering

We filter out any phrase containing digits or punctuation symbols, allowing for a dash “-” inside the phrase, but not at the beginning or at the end. For example, we filter out phrases like *b./m.* “*10 февруари*” (‘*February 10*’), or *b./m.* “*либерал* -” (‘*liberal* -’), or *b./m.* “- *членки*” (‘- *members*’, which is part of *b./m.* “*страни-членки*”, i.e. ‘*member states*’). However, we keep the Macedonian phrases *m.* “*бугарскиот заменик-премиер*” (‘*the Bulgarian vice-premier*’) or *m.* “*анти-корупциска комисија*” (‘*anti-corruption commission*’), where the dash is inside the phrase.

We also filter out any phrase starting or ending with a stopword (e.g. *ги*, *го*, *да*, *до*, *за*, *на*, *но*, *пред*, *се*, *со*, *че*, etc.), but we allow for stopwords inside the phrases. For example, we filter out phrases like *m.* “*разговори со*” (‘*talks with*’), or *b.* “*на България*” (‘*of Bulgaria*’), but we keep *b.* “*комисия за борба с корупцията*” (‘*anti-corruption commission*’), where the stopword *за* is inside the phrase.

The investigation of the lexical features and the syntactic structure of the remaining phrase pairs revealed that some extra filtering might be needed in order to achieve perfect results. While in many cases the process could be automated, in some other cases linguistic analysis would be required, and the boundary between the two is quite fuzzy.

Sample phrases: both correct and wrong can be seen in Table 4.

3 Experiments and Evaluation

3.1 Resources

In our experiments, we make use of a Bulgarian-Macedonian bilingual corpus, consisting of 16,744 aligned sentence pairs: a total of 383,615 Bulgarian and 401,327 Macedonian word tokens. The corpus is part of the multilingual *Balkan South-East Corpus* [22], created at the Linguistic Modeling Department of the Institute for Parallel Processing, Bulgarian Academy of Sciences and has been sentence-aligned using the specialised aligner *LORA* (*Language Objects Raw Analyzer*); see the description in [22] for details. The corpus contains parallel news in nine languages, (Albanian, Bulgarian, Croatian, English, Greek, Macedonian, Romanian, Serbian and Turkish) from *Southeast European Times*[6]; it is both structurally simple and lexically rich, which makes it perfect for the purpose of automatic lexicon extraction.

3.2 Experiments

In our experiments, we applied the five steps of our method to the above sentence-aligned corpus. The cognate extraction step 1, yielded a total of 26,326 distinct cognate pairs. On step 2, we appended them to the corpus twice, and we constructed word alignments using this augmented corpus and IBM model 4. On step 3, we used these word alignments and the alignment template approach in order to extract phrase pairs of length up to seven, obtaining a total of 1,317,635 phrase pairs. The statistical filtering on step 4 reduced this number by about 90% to 137,636. The automatic linguistic filtering on step 5 further reduced the phrase table size to 44,327 phrase pairs.

3.3 Evaluation

3.3.1 Phrase Subgroups

For each of the 44,327 extracted phrase pairs, we calculate a measure of the quality of that phrase, as the minimum of the two conditional probabilities $\Pr(m|b)$ and $\Pr(b|m)$. For the purpose of evaluation, we divided the phrases into ten equal intervals according to this measure, and we evaluated the best five of them: $[0.5; 0.6]$, $(0.6; 0.7]$, $(0.7; 0.8]$, $(0.8; 0.9]$ and $(0.9; 1.0]$. This reduced the total number of phrases we included in our manual evaluation[7] to 16,851.

Since our phrases are of length up to seven, we take into account the phrase length in the manual evaluation and linguistic analysis. We further divide the

[5] We use a maximum phrase length of seven in our experiments.

[6] `http://www.setimes.com`

[7] Our observation is that most of the remaining phrases that we did not include in the manual evaluation, are good as well, e.g. the ones of score 0.05 or more.

phrase pairs into two independent sets: word-to-word translation pairs and multi-word-to-multi-word translation pairs.

3.3.2 Named Entities

News texts are rich in named entities (NEs), which represent an essential part of the information content of the document. This well-known fact has been confirmed by the high percentage, about 40%, of phrases containing NEs in the analysed phrase pairs: see Table 2 for the word-to-word phrase pairs and Figure 1) for the multi-word phrase pairs.

We decided to exclude the NEs from the final evaluations, since, being dynamic lexical elements, they are very unlikely to be of interest as dictionary units.

3.3.3 Coordinated Phrases

As we said above, the linguistic filtering on step 5 allows for stopwords inside the phrases (but not at phrase beginning or ending). In particular, it allows for phrases containing coordinating conjunctions like *b./m. и* (‘*and*’), *b./m. или* (‘*or*’), etc. Since a coordinating conjunction joins two linguistic units, it cannot appear in an initial or in a final position in a legitimate phrase; therefore, we have filtered out all such phrases on step 5 of the algorithm. The phrase-internal coordinating conjunctions however, pose a particular challenge: in some cases they coordinate two independent linguistic units, while in other cases the resulting unit represents an entity with a strong cohesion between its parts, which makes it a non-breakable unit, e.g.

- Justice and Development Party: *b. Партия за справедливост и развитие*, and *m. Партија на правдата и развојот*;
- Central and Eastern Europe: *b. Централна и Източна Европа*, and *m. Централна и Источна Европа*;
- Bosnia and Herzegovina: *b./m. Босна и Херцеговина.*

In our investigation, we found that such cases are relatively rare and therefore we removed all phrases with a coordinating conjunction from the evaluation set. This eliminates many problems with incomplete phrases, e.g. extracting *b. Централна и Източна* instead of *b. Централна и Източна Европа.* In future experiments, we will perform this filtering as part of step 5 of our algorithm.

3.3.4 Word-to-Word Translation Pairs

In our evaluation, we analyse the word-to-word and the phrase-to-phrase translation pairs separately, which reflects the fact that multi-word phrase pairs can be bad not only because of wrong translation, but also because of wrong phrase boundaries. Therefore, while for multi-word phrases we assess both the translation precision and the extraction precision, for word-to-word pairs we analyse the translation precision only.

As can be seen in Table 2, our analysis shows that 40,85% of the word-to-word pairs (i.e. 2,955 out of all 7,235) are named entities, and therefore were excluded from the translation quality evaluation. The remaining 4,280 word-to-word pairs represent correct translations in 99,30% of the cases, as can be seen in Table 3. We believe this high percentage, which is close to the quality of the human translations, is due to two complementary reasons: the reliability of the method and the lexical closeness between the languages.

Interval	Total	NEs	non-NEs	% NEs
(0.9; 1.0]	3,614	1,680	1,934	46.49%
(0.8; 0.9]	307	189	118	61.56%
(0.7; 0.8]	833	393	440	47.18%
(0.6; 0.7]	1,575	520	1,055	33.02%
[0.5; 0.6]	906	173	733	19.09%
Overall	**7,235**	**2,955**	**4,280**	**40.84%**

Table 2: NEs proportion for the word-to-word pairs. *Number and proportion of named entities in the extracted phrases, shown by interval.*

Interval	non-NEs	corr.	wrong	P%
(0.9; 1.0]	1,934	1916	18	99.07%
(0.8; 0.9]	118	116	2	98.31%
(0.7; 0.8]	440	437	3	99.32%
(0.6; 0.7]	1,055	1,051	4	99.62%
[0.5; 0.6]	733	730	3	99.59%
Overall	**4,280**	**4,250**	**30**	**99.30%**

Table 3: Non-NEs translation precision for the word-to-word pairs. *Number of correct and wrong translations for non-NEs, and translation precision in %, shown by interval.*

The analysis of the incorrect translations shows that they stem primarily from lexical gaps and syntactic transformations where a word is translated with a phrase or vice versa. As a result, in some cases a word could be wrongly paired with part of its true phrasal translation (e.g. just one word from the target phrase). For example, there is no corresponding verb in Bulgarian for the Macedonian verb *каменувам* (‘*to throw stones*’), which is therefore translated by a verbal phrase: *b. замервам с камъни.* This causes the generation of a wrong word-to-word translation pair: *m. каменувам – b. замервам.*

Similarly, the corresponding metaphorical expressions *m. покачувам температурите* (lit. ‘*to rise the temperatures*’) and *b. разпалвам страстите* (lit. ‘*to inflame passion*’), give rise to the following wrong word-to-word pair: *m. покачувам – b. разпалвам.*

The semantic transformation with negation shifting gives the paradox pair *m. здраво – b. нездраво*, i.e. ‘*healthy*’ – ‘*unhealthy*’, for the phrases *m. не е здраво* (‘*(it) is not healthy*’) and *b. е нездраво* (‘*(it) is unhealthy*’) .

3.3.5 Multi-word Translation Pairs

We evaluate the multi-word phrases (word length 2–7) according to the following two criteria:

- extraction precision;
- translation precision.

The former criterion is syntactic and checks the structural completeness of the phrase, while the latter is semantic and checks whether the Bulgarian-Macedonian phrase pair represents a correct translation pair.

It is easy to see that the former does not automatically imply the latter. Consider for example the English-Bulgarian phrase pair *e. the shadow of your smile* and *b. тъмнината на твоята усмивка* (‘*the darkness of your smile*’). While this could be a correct extraction, it is not a correct translation, since *the shadow* is translated with *тъмнината* (‘*the darkness*’).

It is also possible to have a correct translation, without a correct extraction. For example, *b. коалиция на местно* (‘*coalition on a local*’) and *m. коалиција на локално* are incorrectly extracted parts of *b. коалиция на местно ниво* (‘*coalition on a local level*’) and *m. коалиција на локално ниво*.

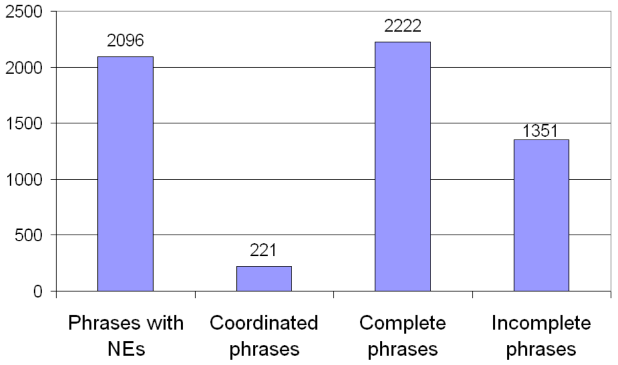

Fig. 1: Multi-word phrases. *Distribution of NEs, coordinated phrases, complete and incomplete phrases in the multi-word phrases: length between two and seven words.*

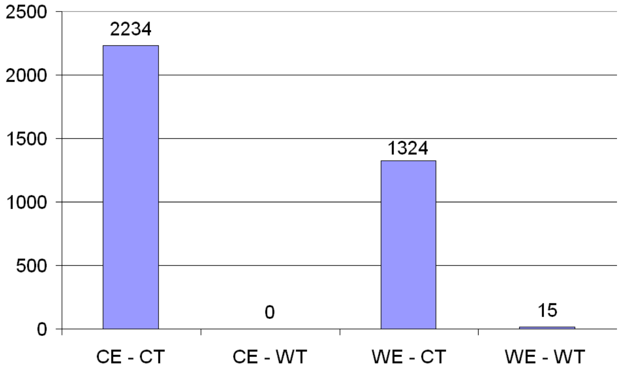

Fig. 2: Complete and incomplete multi-word phrases. *Distribution of the four combinations of values of extraction and translation correctness:* correctly extracted (CE)*,* correctly translated (CT)*,* wrongly extracted (WE) *and* wrongly translated (WT)*.*

As Table 2 shows, 221 of the total 5,890 multi-word phrases (3.75%) contain a coordinating conjunction. As we said in section 3.3.2 above, we have excluded all such phrases from the manual evaluation. Table 2 also shows that 2,096 of the 5,890 multi-word phrase pairs (39.34%) contain an NE. As for the word-to-word translation pairs (see section 3.3.4), we removed these phrases as well. This left us with a total of 3,573 phrase pairs with scores in the interval $[0.5; 1.0]$ to evaluate. As Figure 2 shows, 62,52% of them (i.e. 2,234 out of 3,573, see the CE-CT group on Figure 2) are both correctly extracted and correctly translated, while another 37,06% (i.e. 1,324 out of 3,573, see the WE-CT group on Figure 2) are partially extracted, but still correctly translated. The phrase pairs from the latter group (WE-CT) would have met the requirements for a correct translation if the objective was translation-oriented, but they cannot be treated as dictionary units. Nevertheless, the overall quality of the extracted phrases is very high.

Interestingly, we have no cases of correct extraction, but wrong translation (CE-WT group on Figure 2), which is partially due to the minimum score of 0.5 we impose, and to the good overall quality of the human translations of the text in the original corpus.

Finally, only 0.42% of the phrase pairs (i.e. 15 out of 3,573, see the WE-WT group on Figure 2) are both wrongly extracted and wrongly translated, which means that the overall translation quality is very high.

Table 4 shows sample example phrase pairs of the categories CE-CT, WE-CT and WE-WT with translations into English.

4 Related Work

Many researchers have exploited the intuition that words in two different languages with similar or identical spelling are likely to be translations of each other.

Al-Onaizan&al.’99 [2] create improved Czech-English word alignments using probable cognates extracted with one of the variations of LCSR [15] described in [26]. They tried to constrain the co-occurrences, to seed the parameters of IBM model 1, but their best results were achieved by simply adding the cognates to the training corpus as additional “sentences”. Using a variation of that technique, Kondrak&al.’03 [13] demonstrated improved translation quality for nine European languages.

Nakov&al.’07 [18] propose a method for achieving improved word alignments using the Web as a corpus, a glossary of known word translations (dynamically augmented from the Web using bootstrapping), the vector space model, linguistically motivated weighted minimum edit distance, competitive linking, and the IBM models. Evaluation results on a Bulgarian-Russian corpus show a sizable improvement both in word alignment and in translation quality.

Koehn&Knight’02 [11] describe several techniques for inducing translation lexicons. Starting with unrelated German and English corpora, they look for (1) identical words, (2) cognates, (3) words with similar frequencies, (4) words with similar meanings, and (5) words with similar contexts. This is a bootstrapping process, where new translation pairs are added to the lexicon at each iteration. The method generates word-level correspondences only.

Rapp [23] describes a correlation between the co-occurrences of words that are translations of each other. In particular, he shows that if in a text in one language two words A and B co-occur more often than expected by chance, then in a text in another language the translations of A and B are also likely to co-occur frequently. In later work on the same problem, Rapp [24] represents the context of the target word with four vectors: one for the words immediately preceding the target, another one for the ones immediately following the target, and two more for the words one more word before/after the target.

Fung and Yee [7] extract word-level translations from non-parallel corpora. They count the number of sentence-level co-occurrences of the target word with a fixed set of "seed" words in order to rank the candidates in a vector-space model using different similarity measures, after normalisation and TF.IDF-weighting [25]. The process starts with a small initial set of seed words, which are dynamically augmented as new translation pairs are identified.

Diab & Finch [6] propose a statistical word-level translation model for comparable corpora, which finds a cross-linguistic mapping between the words in the two corpora such that the source language word-level co-occurrences are preserved as closely as possible.

5 Conclusions and Future Work

We have proposed a novel approach for automatic translation lexicon extraction from parallel sentence-aligned bilingual corpora, trying to induce meaningful linguistic units (pairs of words or phrases) that could potentially be included as entries in a bilingual dictionary. We have carefully evaluated the method, thus demonstrating its potential for two very closely-related South-Slavonic languages: Bulgarian and Macedonian. Starting with a parallel sentence-aligned corpus of about 400,000 words for each language, we have built a set of translation pairs consisting of 4,250 word-to-word correspondences (which could be reduced further to approximately 3,850 lemmata) and 3,529 multi-word-to-multi-word correspondences (62.4% of them could be directly included as entries in a dictionary or in a glossary).

The obtained results and the potential further development of the method represent a solid basis for the creation of corpus based bilingual dictionaries, various lexicographical collections, comparative research and typological investigations, statistic measurement of lexical closeness between languages and various multi-lingual applications.

We plan to further improve the method by incorporating named entity recognition, part-of-speech tagging and shallow parsing in the phrase extraction process, which should help to reduce the problems with partial extractions and coordinations. We would also like to try more sophisticated cognate recognition approaches, e.g. the ones described in [18] and [19]. Finally, we plan to try the method for other Balkan language pairs from the *Balkan South-East Corpus* [22].

Acknowledgements. The work of Veno Pacovski is partially supported by the project BIS-21++, FP6 contract no. INCO-CT-2005-016639.

References

[1] A. Agresti. *An Introduction to Categorical Data Analysis.* John Wiley & Sons, 1996.

[2] Y. Al-Onaizan and K. Knight. Translating named entities using monolingual and bilingual resources. In *Proceedings of ACL*, 2001.

[3] S. Bergsma and G. Kondrak. Alignment-based discriminative string similarity. In *Proceedings of ACL'07*, pages 656–663, 2007.

[4] J. A. Bickford and D. Tuggy. Electronic glossary of linguistic terms (with equivalent terms in Spanish). http://www.sil.org/mexico/ling/glosario/E005ai-Glossary.htm, April 2002.

[5] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[6] M. Diab and S. Finch. A statistical word-level translation model for comparable corpora. In *Proceedings of RIAO*, 2000.

[7] P. Fung and L. Y. Yee. An IR approach for translating from nonparallel, comparable texts. In *Proceedings of ACL*, volume 1, pages 414–420, 1998.

[8] H. Johnson, J. Martin, G. Foster, and R. Kuhn. Improving translation quality by discarding most of the phrasetable. In *Proceedings of EMNLP-CoNLL'07*, pages 967–975, 2007.

[9] P. Koehn. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA*, pages 115–124, 2004.

[10] P. Koehn. Europarl: A parallel corpus for evaluation of machine translation. In *Proceedings of the X MT Summit*, 2005.

[11] P. Koehn and K. Knight. Learning a translation lexicon from monolingual corpora. In *Proceedings of ACL workshop on Unsupervised Lexical Acquisition*, pages 9–16, 2002.

[12] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of NAACL*, pages 48–54, 2003.

[13] G. Kondrak, D. Marcu, and K. Knight. Cognates can improve statistical translation models. In *Proceedings of HLT-NAACL 2003 (companion volume)*, pages 44–48, 2003.

[14] G. Mann and D. Yarowsky. Multipath translation lexicon induction via bridge languages. In *Proceedings of NAACL'01*, pages 1–8, 2001.

[15] D. Melamed. Automatic evaluation and uniform filter cascades for inducing N-best translation lexicons. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 184–198, 1995.

[16] D. Melamed. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130, 1999.

[17] I. D. Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, 2000.

[18] P. Nakov, S. Nakov, and E. Paskaleva. Improved word alignments using the Web as a corpus. In *Proceedings of RANLP'07*, 2007.

[19] S. Nakov, P. Nakov, and E. Paskaleva. Cognate or false friend? Ask the Web! In *Proceedings of the RANLP'2007 workshop: Acquisition and management of multilingual lexicons.*, 2007.

[20] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 2003.

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, 2002.

[22] E. Paskaleva. Balkan South–East Corpora Aligned to English. In *Proceedings of the RANLP'2007 workshop: A Common Natural Language Processing Paradigm for Balkan Languages.*, 2007.

[23] R. Rapp. Identifying word translations in non-parallel texts. In *Proceedings of ACL*, pages 320–322, 1995.

[24] R. Rapp. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of ACL*, pages 519–526, 1999.

[25] K. Sparck-Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[26] J. Tiedemann. Automatic construction of weighted string similarity measures. In *Proceedings of EMNLP-VLC*, pages 213–219, 1999.

English Translation	Macedonian Phrase	Bulgarian Phrase
Correct Extraction & Correct Translation		
author	автор	автор
author	автор	автор
the authors	авторите	авторите
the author (female)	авторката	авторката
agent	агент	агент
aggression	агресија	агресия
prisoners	затвореници	затворници
protection	заштита	защита
health (related to)	здравствени	здравословни
associations	здруженија	асоциации
countries	земји	страни
the countries	земјите	страните
slander	клеветата	клеветата
the aggression	агресијата	агресията
the banking sector	банкарскиот сектор	банковия сектор
the bank in the country	банката во земјата	банката в страната
the bank for Europe	банката за Европа	банката за Европа
similar message	слична порака	подобно послание
the next local elections	следните локални избори	следващите местни избори
the freedom of the media	слободата на медиумите	свободата на медиите
case of war	случај на војна	случай на война
bank accounts were frozen	банкарски сметки беа замрзнати	банкови сметки бяха замразени
the requirements of the European commission	барањата на европската комисија	изискванията на европейската комисия
the increased complexity of the expanded union	зголемената сложеност на проширената унија	увеличаващата се сложност на разширения съюз
the measures for strengthening the security	мерките за зајакнување на безбедноста	мерки за засилване на сигурността
European convention for human rights protection	европската конвенција за заштита на човековите права	европейската конвенция за защита на човешките права
candidate for the position of minister of the interior	кандидат за местото министер за внатрешни работи	кандидат за поста вътрешен министър
national service for fighting the organised crime	национална служба за борба против организираниот криминал	национална служба за борба с организираната престъпност
cooperation between the economic chambers of the two countries	соработка помеѓу стопанските комори на двете земји	сътрудничество между търговските камари на двете държави
participation of the Kosovo Serbs in the parliamentary elections	учеството на косовските срби на парламентарните избори	участието на косовските сърби в парламентарните избори
central bank decided not to accept	централна банка одлучи да не го прифати	централна банка реши да не приема
Wrong Extraction & Correct Translation		
forest lost (*forest* is part of *Montenegro*='*black forest*')	гора загуби	гора загуби
forest did not succeed	гора не успеаја	гора не успяха
forest wants	гора сака	гора желае
forest agreed	гора се согласија	гора се споразумяха
two big	две големи	две големи
both ethnic	двете етнички	двете етнически
both Cypriot	двете кипарски	двете кипърски
crimes to avoid justice	злосторства да ја избегнат правдата	престъпления да избегнат правосъдието
the package provides changes according to	пакетот предвидува измени во согласност	пакетът предвижда промени в съответствие
death penalty with life inprisonment	смртната казна со доживотен затвор	смъртното наказание с доживотен затвор
won the elections	победи на изборите	спечели изборите
home ground	домашен терен	у дома
asked the government	побараа од владата	призоваха правителството
demand for foreign currency	побарувачка на девизи	търсене на чуждестранна валута
earthquake	земјотрес	земетресение
efforts	напори	усилия
process	процес	процес
gave a strong	даде силна	даде силна
two bilateral	два билатерални	две двустранни
two Serbian	двајца српски	двама сръбски
Wrong Extraction & Wrong Translation		
	каменуваа	замерваха
	здраво	нездраво
	прекршувања на законите	пункта за нарушения на законите

Table 4: Sample Macedonian-Bulgarian phrase translation pairs.