# BEST FRIENDS OR JUST FAKING IT? CORPUS-BASED EXTRACTION OF SLOVENE-CROATIAN TRANSLATION EQUIVALENTS AND FALSE FRIENDS

Darja FIŠER
University of Ljubljana, Faculty of Arts, Department of Translation

Nikola LJUBEŠIĆ
University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences

*Fišer, D., Ljubešić, N. (2013): Best friends or just faking it? Corpus-based extraction of Slovene-Croatian translation equivalents and false friends. Slovenščina 2.0, 1 (1): 50-77.*
*URL: http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_04.pdf.*

In this paper we present a corpus-based approach to automatic extraction of translation equivalents and false friends for Slovene and Croatian, a pair of closely related languages. While taking advantage of the orthographic similarities between the two languages, the approach relies on a straightforward but powerful assumption of distributional semantics, which stipulates that words with a similar meaning tend to be used in similar contexts in both languages. On the one hand, this phenomenon enables us to quickly generate a Slovene-Croatian bilingual lexicon from minimal knowledge sources, the weakly comparable web corpora. On the other, it can also be used to identify the cognates that only seem similar on the surface but are in fact used to express different concepts in the two languages. The presented approach is language-independent and therefore attractive for natural language processing tasks that often lack the lexical resources and cannot afford to build them by hand, but is also useful in lexicography and language pedagogy where it can be used to highlight the lexical characteristics specific for a given language pair or domain.

**Keywords:** automatic bilingual lexicon extraction, distributional semantics, closely related languages, cognates, false friends

## 1 INTRODUCTION

There is a long tradition of bilingual lexical resources in Slovenia and in Croatia but, not surprisingly, we observe a strong bias towards the major languages, such as English, German and French in both communities. While some bilingual dictionaries including Serbian, Slovakian, Czech, Polish and Russian do exist, they are significantly smaller in size, not updated, and, most importantly, not available in electronic form, especially as complete datasets for research. This lack of resources poses a problem for language learning but also acts as a major inhibitor of human language technologies, such as machine translation applications, the development of which would be extremely welcome also for less mainstream language pairs.

In the past decade or so, researchers have ameliorated the problem by automatically extracting bilingual lexicons from parallel corpora (Och and Ney 2000) but even such corpora exist only for a limited number of language pairs and domains and it is often difficult to build one from scratch. This is why an alternative approach that relies on non-parallel texts in two different languages has become increasingly popular in recent years (Sharoff et al. 2013). Plenty of such corpora exist already, and new ones are much easier to compile, especially from the increasingly rich web data (Baroni and Bernardini 2006; Pomikalek et al. 2009).

The underlying assumption of the non-parallel approach is that the source word and its translation appear in similar contexts (Fung 1998; Rapp 1999), allowing us to identify equivalence pairs by finding the target word with the most similar context vector to the one of the source word that has been extracted from corpora in the respective languages. However, before vector comparison in two different languages can be performed, the features of source context vectors have to be translated into the target language with a seed lexicon. This can be either an existing traditional bilingual dictionary, a bilingual lexicon that has been extracted from a parallel corpus, or a lexicon bootstrapped from comparable corpora.

Since the goal of this paper is to propose a knowledge-light approach to bilingual lexicon extraction for closely related languages, we too extract a seed

lexicon directly from corpora by taking advantage of orthographic similarities between the source and the target language. In addition to identifying translation equivalents by finding word pairs with the most similar context vectors across the two languages, we also show that the inverse is possible, namely the discovery of false friends. Despite being orthographically similar, two words are considered false friends if their context vectors are dissimilar enough.

The rest of the paper is structured as follows: in Section 2 we give an overview of related work. In Section 3 we present the construction of the resources used in the experiment. Section 4 describes the experimental setup and Section 5 reports on the results of automatic and manual evaluation. We conclude the paper with final remarks and ideas for future work.

## 2 RELATED WORK

Even though automatic, corpus-based identification of translation equivalents and detection of false friends are based on the same principles, they are seen as two separate tasks in the computational lexical semantics community. Attempts to automatically extract bilingual lexica from corpora predate false friends identification, and have since become an established research track that is currently being extended to the extraction of translation equivalents for multi-word units (Morin and Daille 2010), domain-specific terminology (Nakao et al. 2010) as well as polysemous vocabulary items (Fišer et al. 2012). Automatic detection of false friends was initially limited to parallel corpora but has been extended to comparable corpora and web snippets (Nakov et al. 2007). To our knowledge, there have been no attempts to augment or refine the extraction of translation equivalents by weeding out false friends, which seems an obvious way to merge both tasks.

The methods described in this section are all applied to non-parallel data. The task is much easier if the corpora used are comparable but with enough data, even weakly comparable corpora suffice. This is why the terms *non-parallel* and *comparable* are used interchangeably with no difference in meaning throughout the paper.

### 2.1 Extraction of translation equivalents

The beginners of bilingual lexicon extraction from non-parallel data are Fung (1998) and Rapp (1999) whose main assumption is that the source word and its translation share similar contexts. The identification of translation equivalents follows a two-step procedure: first, contexts of words are modelled, and then similarity between the source-language and target-language contexts is measured with the help of a dictionary that acts as a bridge between languages. Most approaches represent contexts with context vectors, which are weighted collections of words appearing next to the word in question.

The most commonly used weighing functions are Log-likelihood (Ismail and Manandhar 2010), TF-IDF (Fung 1998) or PMI (Shezaf and Rappoport 2010). Once context vectors have been built for all the words in both languages, the similarity between a source word's context vector and all the context vectors in the target language is computed. The most typical similarity measures are cosine (Fung 1998), Jaccard (Otero and Campos 2005) and Dice (Otero 2007).

In order to be able to compare context vectors across languages, context vector features have to be translated with a seed dictionary. In the event that such a dictionary is not available Koehn and Knight (2002) show that it is possible to obtain a seed lexicon from identical and similarly spelled words which are directly extracted from non-parallel corpora. In this paper, we improve Koehn and Kinght's approach by iterating the calculation of translation equivalents, extending the seed lexicon on every step with additional information, such as cognates and translation equivalents of the most frequent words from the corpus that received a high confidence score. In addition to the iterative expansion of the seed lexicon, we also carry out a final reranking of translation candidates based on cognates clues, similar to the procedure used by Saralegi et al. (2008).

As opposed to Koehn and Knight (2002), we work with much larger corpora and much closer languages, which is why our seed lexicon is substantially

larger, yielding a higher recall and precision of the extracted translation equivalents that consequently results in a more usable resource in a real-world setting. And finally, we are not limiting our experiments only to nouns, but are working with all content words.

### 2.2 Identification of false friends

The approaches to automatically identify false friends fall into two categories: those that only look at orthographic features of the source and the target word, and those that combine orthographic features with the semantic ones. Orthographic approaches typically rely on combinations of a number of orthographic similarity measures and machine learning techniques to classify source and target word pairs to cognates, false friends or unrelated words and evaluate the different combinations against a manually compiled list of legitimate and illegitimate cognates. This has been attempted for English and French (Inkpen et al. 2005; Frunza and Inkpen 2007) as well as for Spanish and Portuguese (Torres and Aluísio 2011).

Most of the approaches that combine orthographic features with the semantic ones have been performed on parallel corpora where word frequency information and alignments at paragraph, sentence as well as word level play a crucial role at singling out false friends, which has been tested on Bulgarian and Russian (Nakov and Nakov 2009). Work on non-parallel data, on the other hand, often treats false friends candidates as search queries, and considers the retrieved web snippets for these queries as contexts that are used to establish the degree of semantic similarity of the given word pair (Nakov et al. 2007). Apart from the web snippets, comparable corpora have also been used to extract and classify pairs of cognates and false friends between English and German, English and Spanish, and French and Spanish (Mitkov et al. 2007). In their work, the traditional distributional approach is compared with the approach of calculating N nearest neighbors for each false friend candidate in the source language, translating the nearest neighbors via a seed lexicon and calculating the set intersection to the N nearest neighbors of the false friend candidate from the target language. A slightly different setting has been investigated by Schulz et al. (2004) who built a medical

domain lexicon from a closely related language pair (Spanish-Portuguese) and used the standard distributional approach to filter out false friends from cognate candidates by catching orthographically most similar but contextually most dissimilar word pairs.

Our work on false friends identification falls in the semantic category, only that instead of harvesting web snippets directly from the web such as Nakov et al. (2007), we use web corpora that were independently built for each language but since the web contains similar text types and covers similar domains, they could be referred to as weakly comparable, not unlike Mitkov et al. (2007). The three main differences between the work we report on in this paper and the related work are:

1. we identify false friends on a language pair with a large lexical overlap;
2. we do not use a precompiled list of positive and negative cognate examples as a starting point but look for all the possible candidates directly in the corpora; and
3. we look for false friends only among homographs (identically spelled words, such as *boja*, which means *buoy* in Slovene but *colour* in Croatian), not among cognates (similarly spelled words, such as the Slovene adjective *bučen (made of pumpkins* and *noisy)* and its Croatian counterpart *bučan (*only *noisy)*).

This enables us to focus on the problem of identifying false friends without having to search for productive patterns for cognates beforehand and introducing noise by automatic cognate identification. By focusing on the problem of finding contextually dissimilar words, we are able to further develop the methods proposed in the existing literature whereas extending the approach to cognates is planned for the future.

## 3 RESOURCES

In this section we present the three types of resources used in this work: the corpora, the seed lexicon and the gold standards. The corpora had already been compiled and linguistically annotated by Ljubešić and Erjavec (2011) but both the seed lexicon and the gold standard for false friends were built for the

experiments reported on in this paper and were derived from the corpora. The gold standard for translation equivalents was obtained from a traditional printed Serbo-Croatian – Slovene dictionary (Jurančič 1989).

### 3.1 Corpora

The contextual information required for the identification of translation equivalents and false friends was extracted from slWaC and hrWac, Slovene and Croatian web corpora that were compiled from the web by crawling the .hr and .si domains (Ljubešić and Erjavec 2011). Since slWaC contains 380 million words and hrWac 1 billion words, vector comparison for extracting translation equivalents based on this amount of data would be computationally too expensive. We therefore custom-built subcorpora by including only the news domains *jutranji.hr* and *delo.si*, which are on-line editions of national daily newspapers with a high circulation and a similar target audience. Since the domains were crawled at approximately the same time, the newspaper articles report on similar events, which is why the subcorpora are not only of the same genre but also quite comparable in terms of content. The documents had already been tokenized, PoS-tagged and lemmatized, resulting in 15.8 million tokens for Slovene and 13.4 million tokens for Croatian. False friends, on the other hand, are a much less frequent phenomenon, which is why we used the entire web corpora for this part of the task.

### 3.2 Seed lexicon

Unlike extensive lexical resources that exist for major languages, no machine-readable dictionary is available for Slovene and Croatian. Having said that, it is also true that they are very close languages, a property that could be used to compensate the lack of dictionary resources. Just as an illustration, Scannell (2007) calculated a 0.74 cosine similarity on distributions of character 3-grams in Slovene and Croatian. A similar result was obtained for Czech and Slovak (0.70) and for Spanish and Portuguese (0.76), whereas English and German, for example, turned out to have significantly less similar distributions of 3-grams (0.34). We therefore decided to take advantage of the

high degree of language similarity and built a seed lexicon from the comparable news corpus by extracting all identical lemmas that were tagged with the same part of speech in both languages.

As Table 1 shows, the seed lexicon contains about 33,500 entries, 77% of which are nouns. Manual evaluation of 100 random entries for each part of speech shows that nouns have the highest precision (88%) and that harmonic mean of precision for all parts of speech in the dictionary is 69%.

| **POS** | **Size** | **Precision** |
|---|---:|---:|
| Nouns | 25,703 | 88% |
| Adjectives | 4,042 | 76% |
| Verbs | 3,315 | 69% |
| Adverbs | 435 | 54% |
| **Total** | **33,495** | **69%** |

**Table 1:** Analysis of the seed lexicon.

The errors we observed in manual evaluation are mostly Croatian words that appeared in the Slovene part of the corpus. As many as 72% of the erroneous nouns belonged to this type of error (e.g. *šećer, baka, tužba*), followed by 66% of the adjectival errors (e.g. *svjetski, iznerviran, rođen*), 63% of wrong adverbs (e.g. *jako, puno, hitno*) and 55% of the erroneous verbs (e.g. *opljačkati, zagustiti, usuditi*). They probably originated from readers' comments that are written in informal language, which often contains Croatian or Serbian expressions. Such errors could be avoided in by a stricter filtering of the corpus.

Most of the rest were spelling, tagging and lemmatization errors. However, we have also come across some false friends that got into the seed lexicon (e.g. noun *rob*, which means *edge* in Slovene but *slave* in Croatian; adjective *složen*, which means *unanimous* in Slovene but *complex* in Croatian; verb *skloniti*, which means *to stoop* in Slovene but *put away* in Croatian). Such errors in the seed lexicon are potentially much more serious because they can create noise in the translation of context vector features, thereby making the comparison of the vectors harder. This is one of the motivations for focusing on the identification of false friends in the second part of this paper.

### 3.2 Gold standards

Gold standards are very important resources because they make automatic evaluation and comparison of the results obtained from different settings faster and more objective. Since we are dealing with two different tasks in this paper, two different gold standards were required. The first one is intended for the evaluation of our approach to extract translation equivalents, and the second one for the evaluation of false friends identification. The translation equivalents gold standard was constructed from 1000 randomly selected entries (618 nouns, 217 adjectives and 165 verbs) taken from the traditional broad-coverage Serbo-Croatian – Slovene dictionary, which contains around 8,100 entries (Jurančič 1989). Although adverbs are included in seed lexicon extensions based on their positive impact on this task, we do not include them in the gold standard for two reasons: first, many tokens tagged as adverbs in the corpus are mistagged other parts of speech, and second, most adverbs in both Slovene and Croatian can be easily generated from adjectives and there is only a small amount of those for which this does not hold, so that they can be considered a closed word class.

The false friends gold standard contains nominal, verbal and adjectival homographs that appeared in the corpora for both languages and were then manually classified into one of the following three categories: false friends, partial false friends and true equivalents. We use the term *true equivalents* to refer to identically spelled words that have the same meaning and usage in both languages (e.g. adjective *bivši*, which means *former* in both languages), and the term *false friends* for identically spelled words which are used to represent different concepts in the two languages (e.g. noun *slovo*, which means *farewell* in Slovene and *letter in the alphabet* in Croatian). *Partial false friends*, then, are identical words that are polysemous and are equivalent in some of the senses but false friends in others (e.g. verb *dražiti*, which can mean either *irritate* or *make more expensive* in Slovene but only *irritate* in Croatian).

Since a realistic distribution of (partial) false friends and true equivalents for Slovene and Croatian is impossible to estimate but it is a fact that false friends

are a relatively rare phenomenon, whatever the language pair, we tried to make the evaluation as objective as possible by including roughly 60% of true equivalents, 20% of false friends and 20% partial false friends in the gold standard, as can be seen in Table 2.

| | **Adjectives** | **Nouns** | **Verbs** |
|---|---|---|---|
| True equivalents | 130 | 131 | 119 |
| Partial false friends | 30 | 41 | 39 |
| False friends | 40 | 41 | 36 |
| **Total** | **200** | **213** | **194** |

**Table 2:** Gold standard for false friends.

## 4 METHODOLOGY AND PROCEDURE

The entire task is illustrated in Table 3 which shows 20 strongest features of context vectors for two Slovene (*priča – witness* and *zgodba – story*) and two Croatian nouns (*svjedok – witness* and *priča – story*) that were computed from their occurrences in the newspaper subcorpora of slWaC and hrWaC.

A cross-comparison of all context vectors shows that the Slovene noun *priča* is the most similar to its Croatian counterpart *svjedok*. Leaving aside the technical details of building and comparing the context vectors in the two languages, which are discussed in detail later in this section, it can be seen from Table 3 that almost half of context vector features are shared between the languages: *key, protected, influence, prosecution, hearing, statement, danger, questioning* and *questioned*. On the other hand, despite being orthographically identical, the use of the word *priča* in Croatian is quite different from Slovene and there is no overlap between their top 20 strongest context features. In fact, *priča*'s highest-ranked counterpart is the Slovene noun *zgodba*, which can be estimated from the eight overlapping context words of the 20 strongest features: *whole, love, life, sad, tell, interesting, different* and *true*. Therefore, throughout the paper, we refer to Slovene-Croatian pairs, such as *priča – svjedok* and *zgodba – priča*, as translation equivalents, while the pair *priča – priča* is an example of false friends.

| SLO *priča* | FEAT. WEIGHTS | CRO *svjedok* | FEAT. WEIGHTS |
|---|---|---|---|
| **zaščiten** | 0.019 | **krunski** | 0.041 |
| **vplivanje** | 0.018 | **zaštićen** | 0.027 |
| **zaslišan** | 0.017 | iskaz | 0.024 |
| **zaslišati** | 0.015 | **utjecaj** | 0.017 |
| **kronski** | 0.013 | pokajnik | 0.014 |
| **zaslišanje** | 0.012 | **tužiteljstvo** | 0.012 |
| vabljen | 0.009 | svojstvo | 0.010 |
| anonimen | 0.008 | utjecati | 0.010 |
| poročen | 0.008 | **saslušanje** | 0.010 |
| seznam | 0.007 | **izjava** | 0.009 |
| **izjava** | 0.007 | status | 0.009 |
| izpoved | 0.007 | saslušanih | 0.009 |
| **tožilstvo** | 0.007 | ispitati | 0.009 |
| **nevarnost** | 0.006 | optužba | 0.009 |
| pripovedovanje | 0.005 | utjecanje | 0.007 |
| dogodek | 0.005 | **opasnost** | 0.007 |
| **status** | 0.005 | obrana | 0.007 |
| mogetov | 0.005 | saslušati | 0.006 |
| navedba | 0.005 | istraga | 0.006 |
| osivnika | 0.004 | **ispitivanje** | 0.005 |
| pričanje | 0.004 | **ispitan** | 0.0053 |
| **SLO *zgodba*** | **FEAT. WEIGHTS** | **CRO *priča*** | **FEAT. WEIGHTS** |
| uspeh | 0.015 | **cijel** | 0.015 |
| **ljubezenski** | 0.009 | **ljubavan** | 0.007 |
| plat | 0.007 | kuloarski | 0.005 |
| **pripovedovati** | 0.006 | **životan** | 0.005 |
| tragičen | 0.005 | kružiti | 0.004 |
| tajkunski | 0.005 | kratak | 0.003 |
| **življenjski** | 0.004 | akter | 0.003 |
| **žalosten** | 0.004 | **tužan** | 0.002 |
| **celoten** | 0.004 | nastavak | 0.002 |
| podoben | 0.004 | božićan | 0.002 |
| ponoviti | 0.004 | dio | 0.002 |
| **resničen** | 0.004 | poseban | 0.002 |
| nauk | 0.003 | **ispričati** | 0.002 |
| izmišljen | 0.003 | **zanimljiv** | 0.002 |
| **drugačen** | 0.003 | verzija | 0.002 |
| epilog | 0.003 | pozadinski | 0.002 |
| uspešen | 0.003 | ispričan | 0.002 |
| narnija | 0.002 | **pričati** | 0.002 |
| **zanimiv** | 0.002 | **drukčiji** | 0.002 |
| patria | 0.002 | **istinit** | 0.002 |
| **cel** | 0.002 | junak | 0.002 |

**Table 3:** An illustrative example of corpus-based identification of translation equivalents and false friends (the overlapping features in both languages are printed in bold).

### 4.1 Extracting translation equivalents

In the first part of the experiment, our task was to extract a bilingual lexicon from a comparable corpus. We use best-performing settings for building and comparing context vectors from our previous research (see Ljubešić et al. 2011). We built context vectors for all content words in each language with a minimum frequency of 50 occurrences in the corpus. The co-occurrence window was 7 content words with encoded position of context words in that window, and Log-likelihood as vector association measure. Vector features were then translated with the seed lexicon. The seed lexicon was automatically compiled from slWaC and hrWac and contains words from the corpus which are identical in both languages. After that Jensen-Shannon divergence was applied as the vector similarity measure.

In order to improve the results, we experimented with the following extensions of original procedure:

1. extending the seed lexicon with contextually similar cognates; and
2. extending the seed lexicon with first translations of the most frequent words.

Cognates were calculated with BI-SIM, the longest common subsequence of bigrams with a space prefix added to the beginning of each word in order to punish the differences at the beginning of the words (Kondrak and Dorr 2004). The threshold for cognates was empirically set to 0.7 (cf. Ljubešić et al. 2011). Twenty top-ranking translations were taken into account and were analyzed for cognate clues in that order. If a translation equivalent that meets the cognate threshold was found, we added that pair to the seed lexicon. If the seed lexicon already contained a translation for a cognate we identified with this procedure, we replaced the existing dictionary entry with the new identified cognate pair as this setting produced best results in our previous work (Ljubešić et al. 2011).

For the extension of the seed lexicon with first translations of the most frequent words we only took into account the first translation candidates for words that appear at least 200 times in the source corpus. If the seed lexicon

already contained an entry we were able to translate with this procedure, we again replaced the old pair with the new one.

Finally, we reranked the translation candidates of all content words obtained with this procedure by taking into account cognate clues among the candidates. The source word was compared by the previously described BI-SIM function with each of the top ten translation candidates. Two lists were formed, one with words that meet the 0.7 cognate threshold criterion and another one with the words that do not. Apart from that, the order of the words in the lists was unchanged. In the end, the lists were combined by putting the cognate list of translation equivalents in front of the non-cognate list.

### 4.2 Identifying false friends

In the second part of the experiment, our task was to identify false friends from the comparable corpus. Since false friends are a rather rare phenomenon, we did not use the small newspaper subcorpora as in previous experiments, but web corpora in their full size of 380 million words for Slovene and 1.2 billion words for Croatian. Even though false friends can also be found among cognates, we only looked for them among homographs in this experiment. By leaving aside the problem of identifying cognate candidates that could at the same time be false friends, we were able to focus completely on the task at hand – identifying words that are contextually, and therefore semantically, distant enough. It is our belief that there is no difference in the semantic similarity distributions between the group of orthographically identical and orthographically similar false friends, so the same methodology as we propose in this paper could be applied to cognates.

In the web corpora we have identified 8,491 nominal, verbal and adjectival lemmata that pass the 50 occurrences threshold and are orthographically identical in both languages. The gold standard for false friend identification is based on this list and contains 607 entries. We built co-occurence vectors for those entries from the comparable corpora in a similar manner as in the task of extracting translation equivalents. We used content words as features, a 7-

word window, TF-IDF for weighting features and we calculated context similarity with the Dice similarity measure. In our initial experiments these methods have proven to produce best results on this task.

When identifying false friends, we took into account an additional source of information, namely the frequency of the words in each corpus. We assumed that if two identical words have a high discrepancy in frequency between the two languages, this could be a cue that those words do not represent the same meaning. We represented the difference in frequency by calculating Pointwise mutual information.

## 5 EVALUATION AND DISCUSSION OF THE RESULTS

In this section we report on the results of automatic evaluation for both tasks. In order to get a more qualitative insight into the results of the translation equivalents we extracted, we also performed a manual evaluation on a sample of the obtained equivalence pairs.

### 5.1 Evaluation of translation equivalents extraction

#### 5.1.1 AUTOMATIC EVALUATION

In automatic evaluation of the extracted translation equivalents, we measured precision by calculating Mean reciprocal rank (Vorhees 2001, MRR) on the ten top-ranking translation candidates. In this experimental setup, recall for nouns was always 45% because we always found translations for 278 of the 618 nouns from the gold standard that satisfied the frequency criterion (50) in the source corpus and had at least one translation in the target corpus that met the same frequency criterion. For the same reason for other parts of speech recall was also constant: 42% for adjectives and 56.4% for verbs. Overall recall was 46.2%. The baseline precision used for evaluating seed lexicon extensions of 0.592 was calculated by translating features in context vectors of nouns, verbs and adjectives with the seed lexicon of homographs using the settings described in Section 4.1. Baseline precision for individual parts of speech was 0.605 for nouns, 0.579 for verbs and 0.566 for adjectives.

The extended seed lexicon with cognates and first translation candidates contains 2,303 new entries, almost half of which are nouns. The total size of the extended lexicon is therefore 35,798 entries. Precision achieved with the extended seed lexicon was 0.731 (a 0.146 increase).

Table 4 shows the baseline results for all parts of speech, the results obtained by using the extended seed lexicon, and the results of cognate-based reranking of the final translation candidates. From the start, the easiest translation task was that of nouns, followed by verbs while adjectives seem to be the hardest to translate correctly. The biggest gain by extending the seed lexicon has been achieved on nouns (0.163) while verbs and adjectives have experienced a smaller improvement (0.079 and 0.039, respectively). When the results were evaluated on all parts of speech together, the translation results were considerably better than the baseline (0.121 gain), which is still worse than the best-performing nouns, probably due to the noise introduced by verbs and adjectives.

| POS | Baseline | Extended | Reranking |
|---|---|---|---|
| Nouns | **0.605** | **0.768** | **0.848** |
| Adjectives | 0.566 | 0.605 | 0.698 |
| Verbs | 0.579 | 0.658 | 0.735 |
| **Total** | 0.592 | 0.713 | **0.797** |

**Table 4:** Automatic evaluation of translation extraction per PoS with reranking.

Reranking the translation candidates with cognate clues helped all parts of speech, improving the harmonic mean of all precisions by 0.083. Reranking worked particularly well with adjectives (15.4%), probably because of the regularity of patterns for forming adjectives in both languages. Nouns and verbs have experienced a similar precision boost (around 11%).

Regarding the final results, the best score has been achieved for nouns with a total increase in precision, which amounts to 24%. Although adjectives have experienced the biggest boost by reranking, their extraction precision remained the lowest. The reason for that could lie in the possibly largest context heterogenity because of their modifying function. The overall improvement of the results for all parts of speech was 20.5%.

These figures confirmed the positive impact of exploiting language similarity on knowledge-light extraction of bilingual lexicons from comparable corpora for closely related languages. Last but not least, the described method results in a fully automatically created resource the quality of which already makes it useful for practical tasks.

#### 5.1.2 MANUAL EVALUATION

For a more qualitative insight into the results we also performed manual evaluation on a sample of 100 random translation equivalents of the best-performing settings, i.e with the extended seed lexicon and reranking of the translation candidates. The evaluation shows that 88 of the 100 word pairs we checked contained a correct translation among the ten top-ranking translation candidates. 64 of those were found in the first position and 24 in the remaining nine positions, which is a significant improvement compared to the baseline (0.597). What is more, many lists of ten top-ranking translation candidates contained not one but several correct translation variants. Also, as many as 59 of correct translation candidates were cognates and 41 of them even appeared in the first position, suggesting that a final reranking of translation candidates based on cognate clues is highly beneficial.

Table 5 shows some examples of ten top-ranked Slovene translation candidates for some Croatian nouns. In a number of cases the correct translation equivalent is ranked the highest (e.g. *lanac – veriga (chain)*). There are even cases where more than one correct translation equivalent is found (e.g. *protivnik – nasprotnik, tekmec (opponent)*). At other times the correct translation is found among the candidates but is not ranked the highest (e.g. *ušteda – prihranek (savings)*). In most of these cases the highest ranked erroneous candidate is semantically closely related to the correct translation (e.g. *ušteda (savings) – poraba* (expenditure)*, which is the antonym of the correct translation). Occasionally, the correct translation is not found in the list of ten top ranked candidates but the translation candidates are semantically related to the correct translation (e.g. *travnjak (meadov) – igrišče (playing field), parket (parquet), zelenica (lawn), led (ice), navijač (fan), tekma (match), moštvo (team), gol (goal), zadetek*

*(score), vratar (goalkeeper)*).

| Source | Translation candidates |
|---|---|
| lanac | **veriga**, marža, polica, center, lokal, znamka, trgovina, trgovec, proizvajalec, središče |
| svečanost | **slovesnost**, prireditev, proslava, shod, srečanje, concert, delavnica, obletnica, festival, večerja |
| preokret | **preobrat**, zasuk, presenečenje, čudež, neznanka, polom, škoda, blamaža, napaka, uspeh |
| šaka | **pest**, rep, glava, noga, roka, maslo, streha, vrat, strela, trebuh |
| protivnik | **nasprotnik**, **tekmec**, opcija, elita, igrica, zapornik, ambicija, prizorišče, scena, tekmica |
| ušteda | poraba, **prihranek**, znesek, vsota, povečanje, zmanjšanje, porabnik, količina, izboljšava, zniževanje |
| izražaj | poštev, plano, **izraz**, vrsta, spoznanje, zamuda, sonce, misel, zastoj, streznitev |
| dopuna | zakon, novela, **dopolnitev**, sprejetje, osnutek, člen, določba, sprememba, predlog, uveljavitev |
| naljepnica | motor, uniforma, hladilnik, tabla, transparent, črpalka, pogon, plakat, **nalepka**, kolona |
| travnjak | igrišče, parket, zelenica, led, navijač, tekma, moštvo, gol, zadetek, vratar |

**Table 5:** Examples of ten top-ranked Slovene translation candidates for some Croatian nouns (the correct translation equivalents are printed in bold).

### 5.2 Evaluation of false friends identification

In the translation equivalents extraction task, we obtained a ranked list of translation candidates for each word in the source language. In the false friends identification task, however, we obtained a single ranked list where pairs of identical words were sorted in reverse order according to their context similarity.

Since in this task, unlike in the task of extracting translation equivalents, the ranked list contains a number of entries that are actual false friends, we were not able to use MRR for evaluation since it records only the position of the first hit. We therefore decided to use average precision (AP), which is the measure regularly used in the area of information retrieval to evaluate a ranked list of documents as a result of a query. Average precision is the

average of all precision values obtained for the set of top k words that exist after each false friend is located in the ranked list.

Since the gold standard consists of three categories (true equivalents, partial false friends and false friends), we considered three variants of the gold standard:

1. only full false friends are considered false friends;
2. false friends receive weight 1 while partial false friends receive weight 0.5; and
3. both false friends and partial false friends receive weight 1, i.e. both are considered false friends.

Taking into account all the variants of the gold standard, we performed three evaluations on the result of each experiment and calculated the final evaluation measure as harmonic mean of the three evaluation results. As a baseline for this research we used random ordering of false friend candidates. Our first experiment, after calculating baseline performance, was focused on calculating the ranked list with Pointwise mutual information that uses only frequencies of words in both languages. The results of the performance of the specific settings are given in Table 6. It is interesting to see how highly informative just word frequency is, improving the baseline by 29 points.

| Method | AP |
|---|---|
| random baseline | 0.275 |
| pmi (frequencies only) | 0.563 |
| dice, tfidf | 0.637 |
| dice, tfidf>0.01 | 0.692 |
| 0.25*pmi+0.75*dice, tfidf>0.01 | **0.720** |

**Table 6:** Automatic evaluation of false friends identification using four different settings.

Next we performed an experiment with the standard context vector method, using previously confirmed best-performing settings: content words as features, TF-IDF for weighting features and the Dice similarity measure. The increase in performance with respect to using plain frequencies trough PMI

was moderately high (7 points), which stresses once again how well the approach using just the frequency information performs.

In our recent experiments in extracting translation candidates (Appidianaki et al. 2011) we noticed that modifying the TF-IDF weighting by discarding all low feature weights improved the results substantially. We therefore introduced a new weighting scheme called TF-IDF>0.01, which discards feature weights on or below the 0.01 threshold. The improvement obtained by the modified weighting scheme was quite high with a gain of 6 points.

Since the PMI method and the context vector method use completely independent information sources, it seemed natural to try combining the results of both. We decided to perform a simple linear combination of word rankings and experimented with the coefficients of the linear combination. Best results were obtained when contextual information was given greater importance, precisely three times more, than frequency information. By performing the linear combination we gained 3 additional points.

The evaluation scores obtained by calculating the harmonic mean of the results on all three gold standard variations have a very high correlation (>0.99) with more than ten data points for each gold standard. This fact shows that, for this task, it is not crucial how one represents the ternary classified data from the gold standard. It is important to note that all optimizations performed were not using held out data and that for our future work we plan a more formal optimization and evaluation, focused just on the specific problem of false friend identification.

Partial results of the false friends identification procedure is given in Table 7, which contains a list of twenty top-ranked and twenty bottom-ranked false friends candidates according to their context (dis)similarity. As many as 19 (95%) of the 20 top-ranked candidates are genuine false friends, which are also in the gold standard. Most of the top-ranked false friends are adjectives (50%), followed by verbs (30%) and nouns (20%). The bottom of the list, on the other hand, contains contextually the most similar words, all of which are legitimate equivalents.

| Ranked FF candidates | | | Eng. translation | |
|---|---|---|---|---|
| **FF candidate** | **POS** | **Weight** | **Croatian** | **Slovene** |
| **priljubljen** | A | 21.4 | close together | popular |
| **opasan** | A | 29.8 | dangerous | girt |
| **zarobiti** | V | 33.7 | enslave | hem |
| **pogoditi** | V | 50.4 | hit, agree | agree |
| **zoran** | A | 78.5 | obvious | ploughed |
| **naglašen** | A | 86.2 | stressed | stressed, accented |
| **skriviti** | V | 91.3 | commit | bend |
| **otopiti** | V | 130.0 | melt | make blunt, become numb |
| **zarobljen** | A | 153.1 | enslaved | hemmed |
| čuvan | A | 158.5 | guarded | guarded |
| **valjan** | A | 159.2 | valid, rolled | rolled |
| **ustrojen** | A | 159.8 | constituted | tanned (leather) |
| **zarediti** | V | 163.5 | turn into a priest | infest |
| **boja** | N | 169.1 | colour | buoy |
| **meta** | N | 169.4 | target | mint |
| **približan** | A | 171.8 | approximate | moved closer |
| **žaljenje** | N | 172.1 | grief | insult |
| **iskanje** | N | 175.6 | request | search |
| **razglašen** | A | 184.0 | announced | announced, out of tune |
| **stradati** | V | 189.6 | get hurt | starve |
| ... | ... | ... | ... | ... |
| predlagati | V | 2557.3 | suggest | suggest |
| magistrirati | V | 2558.4 | receive MA degree | receive MA degree |
| narezati | V | 2558.7 | cut | cut |
| esej | N | 2566.5 | essay | essay |
| hrana | N | 2567.1 | food | food |
| jagoda | N | 2580.6 | strawberry | strawberry |
| akcijski | A | 2593.9 | action | action |
| animacijski | A | 2595.4 | animation | animation |
| dizelski | A | 2605.3 | diesel | diesel |
| klarinet | N | 2608.4 | clarinet | clarinet |
| komedija | N | 2636.7 | comedy | comedy |
| naslikati | V | 2645.7 | paint | paint |
| animiran | A | 2646.9 | animated | animated |
| ljubiti | V | 2651.1 | kiss | love |
| kemija | N | 2653.8 | chemistry | chemistry |
| bazenski | A | 2666.3 | pool | pool |
| film | N | 2699.1 | film | film |
| doktorski | A | 2715.3 | doctoral | doctoral |
| junak | N | 2748.8 | hero | hero |
| hokej | N | 2799.7 | hockey | hockey |

**Table 7:** Twenty top-ranked and twenty bottom-ranked false friends candidates (genuine false friends are printed in bold).

### 5.3 Evaluating the impact of removing the identified false friends from the seed lexicon on the quality of translation equivalents extraction

A final experiment was performed by merging the results obtained from false friend identification and the baseline experiment on translation equivalent extraction. The aim of the experiment was to see if discarding the strongest false friend candidates from the initial seed lexicon of homographs could improve the translation candidate extraction results. The results are presented in Table 8.

| # FF | FF standard | FF improved |
|---|---|---|
| 100 | **0.605** | 0.603 |
| 500 | 0.604 | **0.615** |
| 1000 | 0.602 | 0.611 |
| 2000 | **0.605** | 0.605 |
| | **Baseline** | **0.605** |

**Table 8:** The impact of discarding false friends from the seed lexicon on the extraction of translation equivalents.

We removed entries from the seed lexicon which occurred on top 100, 500, 1000 or 2000 positions in the ranked list of false friend candidates. We considered two false friends candidate lists. The *Standard FF ranked list* is the one obtained by the standard distributional method while the *Improved FF* ranked list is the one obtained through the best performing method, i.e. by combining the standard distributional method with an improved weighting scheme and frequency information through PMI.

By using the *Standard FF list* we did not notice any improvement regardless of the number of the first candidates removed from the seed lexicon. By using the *Improved FF list* we did notice a moderate improvement. We consider these results pointing at two facts:

1. false friends are a rare phenomenon and their impact on the task of translation equivalent extraction is limited, especially taking into account the large size of the seed lexicon of 33,000 entries; and
2. the *Improved FF* method outperforms the *Standard FF* method.

We can conclude that the identification of false friends can be beneficial for the task of translation equivalents extraction from comparable corpora provided that the list of false friends is of good quality and the number of false friends in the seed lexicon is substantial. However, identifying false friends between two languages for purposes of training translators and second language acquisition is of great importance just as well.

## 6 CONCLUSION

In this paper we presented a knowledge-light approach to extract translation equivalents and false friends from non-parallel corpora of similar languages. The extraction of translation equivalents outperformed related approaches in terms of precision (0.592 vs. 0.797, with nouns reaching as high as 0.848). Unlike most related approaches it deals with all content words, and enriches the seed lexicon used for translating context vectors from the results of the translation procedure itself.

Although less mature at this stage, our corpus-based attempts to identify false friends have proven to be successful as well, especially when combining context-based and frequency-based feature comparisons, resulting in 0.720 average precision. When these best-performing settings were used to eliminate false friends from the automatically generated seed lexicon, they achieved a very limited improvement of the results in the translation equivalence extraction task, but this should not decrease its importance for both language teaching and more fine-grained natural language processing tasks.

The proposed approach is directly applicable to a number of other similar language pairs for which there is a lack of bilingual lexica. This makes it an attractive starting point for a number of natural language processing, language teaching as well as lexicographic tasks.

The biggest obstacles in false friends evaluation were the lack of an authoritative and comprehensive gold standard, and the lack of information on the frequency, distribution and nature of false friends with respect to legitimate homograph/cognate pairs between two related languages. As a

consequence, the construction of a high quality gold standard is far from trivial, which is why ours, which currently contains 607 true equivalents, partial false friends and false friends, will have to be improved in the future by taking into account inter-annotator agreement and modifying the distribution of false friends and identical words to make it more realistic. In addition, we should pay more attention to some regularities between false friends we have come across, such as the difference between completely accidental lexical overlap (e.g. noun *meta* which means *mint* in Slovene but *target* in Croatian, or noun *sat*, which means *honeycomb* in Slovene but *hour* in Croatian) and etymologically related word pairs the usage of which has diverged over time (e.g. verb *važiti*, which means *to show off* in Slovene and *to be valid* in Croatian, or verb *stradati*, which means *to starve* in Slovene and *to get hurt* in Croatian). We have also observed some regularities of morpho-semantically motivated prefixes and suffixes (e.g. the ending *–en* in Croatian adjectives which often corresponds to participial adjectives in Slovene, such as in *neodgovoren*, which means *unanswered* in Croatian but *irresponsible* in Slovene; the correct Slovene equivalent for unanswered would be *neodgovorjen*). Apart from improving the gold standard, we also wish to fine-tune the proposed ranking function for false friends by assigning different, PoS-specific weights to context features.

## BIBLIOGRAPHY

Al-Onaizan, Y., and Knight, K. (2002): Translating Named Entities Using Monolingual and Bilingual Resources. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL'02*: 400–408. *Philadelphia*.

Apidianaki, M., Ljubešić, N., and Fišer, D. (2012): Disambiguating vectors for Bilingual Lexicon Extraction from Comparable Corpora. In: *Proceedings of the 15th International Multiconference Information Society, IS-LTC'12*: 10–15. Ljubljana.

Baroni, M., and Bernardini, S. (2006): *Wacky! Working Papers on the Web as Corpus*. Bologna: GEDIT.

Fišer, D., Ljubešić, N., and Kubelka, O. (2012): Addressing Polysemy in Bilingual Lexicon Extraction from Comparable Corpora. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC'12*: 3031–3035. Istanbul.

Fišer, D., Ljubešić, N., Vintar, Š., and Pollak, S. (2011): Building and Using Comparable Corpora for Domain-Specific Bilingual Lexicon Extraction. In: *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, BUCC'11*: 19–26. Portland.

Frunza, O., and Inkpen, D. (2007): A Tool for Detecting French-English Cognates and False Friends. In: *Proceedings of the 14th conference Traitement Automatique des Langues Naturelles, TALN'07*, Toulouse.

Fung, P. (1998): A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Nonparallel Corpora. In: *Proceedings of the Third Conference of the Association for Machine Translation in the Americas,* AMTA'98: 11–17. Langhorne.

Inkpen, D., Frunza, O., and Kondrak, G. (2005): Automatic Identification of Cognates and False Friends in French and English. In: *Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing, RANLP'05*: 251–257. Borovets.

Ismail, A., Manandhar, S. (2010): Bilingual Lexicon Extraction from Comparable Corpora Using In-domain Terms. In: *Proceedings of the 23rd International Conference on Computational Linguistics, COLING'10*: 481–489. Beijing.

Jurančič, J. (1989): *Slovensko-srbohrvaški slovar*. Ljubljana: Državna založba Slovenije.

Koehn, P., and Knight, K. (2002): Learning a Translation Lexicon from Monolingual Corpora. In: *Proceedings of the ACL'02 workshop on Unsupervised lexical acquisition, ULA'02*: 9–16. Philadelphia.

Kondrak, G., and Dorr, B. J. (2004): Identification of Confusable Drug Names: A New Approach and Evaluation Methodology. In: *Proceedings of the 20th international conference on Computational Linguistics, COLING'04*, Geneva.

Ljubešić, N., and Erjavec, T. (2011): Compiling Web Corpora for Croatian and Slovene. In: *Proceedings of the Third International Workshop on Balto-Slavonic Natural Language Processing, BSNLP'11*: 395–402. Plzeň.

Ljubešić, N., Fišer, D., Vintar, Š., and Pollak, S. (2011): Bilingual Lexicon Extraction from Comparable Corpora: A Comparative Study. In: *Proceedings of the 1st International Workshop on Lexical Resources, WOLER'11*: 49–54. Ljubljana.

Markó, K., Schulz, S., and Hahn, U. (2005): Multilingual Lexical Acquisition by Bootstrapping Cognate Seed Lexicons. In: *Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing, RANLP'05*: 301–307. Borovets.

Mitkov, R. Pekar, V., Blagoev, D., and Mulloni, A. (2007): Methods for Extracting and Classifying Pairs of Cognates and False Friends. *Machine Translation*, 21 (1): 29–53.

Morin, E., Daille, B. (2010): Compositionality and Lexical Alignment of Multi-Word Terms. *Language Resources and Evaluation*, 44 (1/2): 79–95.

Nakao, Y., Goeuriot, L., and Daille, B. (2010): Multilingual Modalities for Specialized Languages. *Terminology*, 16 (1): 51–76.

Nakov, S., and Nakov, P. (2009): Unsupervised Extraction of False Friends from Parallel Bi-Texts Using the Web as a Corpus. In: *Proceedings of the 6th International Conference on Recent Advances in Natural Language Processing,* RANLP'09: 292–298. Borovets.

Nakov, S., Nakov, P., and Paskaleva, E. (2007): Cognate or False Friend? Ask the Web! In: *Proceedings of the 5th International Conference on Recent Advances in Natural Language Processing, RANLP'07*, Borovets.

Och, F. J., and Ney, H. (2000): Improved Statistical Alignment Models. In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, ACL'00*: 440–447. Hong Kong.

Otero, P. G. (2007): Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In: *Proceedings of Machine Translation SUMMIT XI, MTS'07*: 191–198. Copenhagen.

Otero, P. G., and Campos J. R. P. (2005): An Approach to Acquire Word Translations from Non-parallel Texts. In: *Proceedings of the 12th Portuguese Conference on Artificial Intelligence, EPIA'05*: 600–610. Aveiro.

Pomikalek, J., Rychly, P., and Kilgarriff, A. (2009): Scaling to Billion-plus Word Corpora. *Advances in Computational Linguistics: Special Issue of Research in Computing Science*, 41 (1): 3–14.

Rapp, R. (1999): Automatic Identification of Word Translations from Unrelated English and German Corpora. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics, ACL '99*: 519–526. Stroudsburg.

Saralegi, X., San Vicente, I., and Gurrutxaga, A. (2008): Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In: *Proceedings of the 4th Workshop on Building and Using Comparable Corpora, BUCC'08*: Portland.

Scannell, K. P. (2007): Language Similarity. Dostopno prek: http://borel.slu.edu/crubadan/table.html.

Schulz, S., Markó, K., Sbrissia, E., Nohama, P., and Hahn, U. (2004): Cognate Mapping – A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon. In: *Proceedings of the 20th International Conference on Computational Linguistics, COLING'04*: 813–819. Geneva.

Shao, L., and Ng, H. T. (2004): Mining New Word Translations from Comparable Corpora. In: *Proceedings of the 20th International Conference on Computational Linguistics, COLING'04*, Geneva.

Sharoff, S., Zweigenbaum, P., and Fung, P. (2013): *BUCC: Building and Using Comparable Corpora*. Berlin and Heidelberg. Springer.

Shezaf, D., and Rappoport, A. (2010): Bilingual Lexicon Generation Using Non-Aligned Signatures. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL'10*: 98–107. Uppsala.

Torres, L. S., and Aluísio, S. M. (2011): Using Machine Learning Methods to Avoid the Pitfall of Cognates and False Friends in Spanish-Portuguese Word Pairs. In: *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, STIL'11*: 67–76. Cuiaba.

Vorhees, E. M. (1999): TREC-8 Question Answering Track Report. In: *Proceedings of the Eighth Text REtrieval Conference, TREC-8*: 77–82. Gaithersburg.

# NAJBOLJŠI ALI LAŽNI PRIJATELJI? LUŠČENJE SLOVENSKO-HRVAŠKIH PREVODNIH USTREZNIC IN LAŽNIH PRIJATELJEV IZ KORPUSOV

V prispevku predstavimo korpusni pristop k samodejnemu luščenju prevodnih ustreznic in lažnih prijateljev med slovenščino in hrvaščino. Pristop izkorišča ortografske podobnosti med jezikoma in temelji na predpostavki distribucijske semantike, ki se glasi, da govorci obeh jezikov besede s podobnim pomenom uporabljamo v podobnih kontekstih. To nam po eni strani omogoča hitro izgradnjo slovensko-hrvaškega dvojezičnega leksikona, za katero razen primerljivih spletnih korpusov ne potrebujemo nobenega drugega vira znanja. Po drugi strani pa lahko na podlagi iste predpostavke s pomočjo korpusnih podatkov prepoznamo tiste sorodnice, ki so si podobne zgolj površinsko, leksikalizirajo pa različne pojme in se zato tudi različno uporabljajo. Predstavljen pristop je jezikovno neodvisen, zaradi česar je privlačen za številna področja računalniške obdelave naravnega jezika, kjer še vedno vlada veliko pomanjkanje leksikalnih virov, njihove ročne izdelave pa si ne moremo privoščiti. Pristop je prav tako zelo koristen v leksikografiji in za poučevanje tujih jezikov, saj nam pomaga osvetliti leksikalne posebnosti za določen jezikovni par oziroma strokovno področje.

**Ključne besede:** avtomatsko luščenje dvojezičnega leksikona, distribucijska semantika, sorodni jeziki, sorodnice, lažni prijatelji