An Online Dictionary Browser for Automatically Generated Bilingual Dictionaries[1]

Enikő Héja & Dávid Takács

Keywords: *parallel corpus, proto-dictionary, dictionary query system, semantic relations*.

Abstract

The objective of this paper is to demonstrate that corpus-driven bilingual dictionaries generated fully by automatic means are suitable for human use.

Previous experiments have proven that bilingual resources can be created by applying word alignment on parallel corpora and such resources are useful for bilingual dictionary compilation purposes. Moreover, the corpus-driven nature of the method yields several advantages over more traditional approaches. Most importantly, the exploitation of parallel corpora decreases the reliance on human intuition during dictionary building. However, the proposed technique has to face some difficulties, as well. First, the scarce availability of parallel texts for medium density languages imposes limitations on the size of the resulting dictionary. Secondly, the resulting bilingual resource is not completely clean: that is, wrong translation candidates are also included in the dictionary. In fact, there is a tight correlation between the proportion of wrong candidates and the size of the resulting resource.

Our objective is to design and implement a dictionary a query system that is apt to exploit the additional benefits of the dictionary building method and overcome the disadvantages of it.

1. Introduction

The objective of this research has been to investigate to what extent LT methods are capable of supporting the creation of bilingual dictionaries. Need for such dictionaries shows up specifically in the case of lesser used languages where it does not pay off for publishers to invest into the production of dictionaries due to the low demand. The targeted size of the dictionaries is between 15,000 and 25,000 entries. Since the completely automatic generation of clean bilingual resources is not possible according to the state of the art, we have decided to provide lexicographers with bilingual resources that can facilitate their work. These kind of lexical resources will be referred to as *proto-dictionaries* henceforward. In addition, a dictionary query system has been designed to make the proto-dictionaries available online.

After investigating some alternative approaches, for example hub-and-spoke model (Martin, 2007), alignment of WordNets, we have decided to use word alignment on parallel corpora to generate proto-dictionaries. Previous experiments (Héja, 2010) have proven that word alignment is not only able to help the dictionary creation process itself, but the proposed technique also yields some definite advantages over more traditional approaches. The main motivation behind our choice was that the corpus-driven nature of the method decreases the reliance on human intuition during lexicographic work. Although the careful investigation of large monolingual corpora might have the same effect, being tedious and time-consuming it is not affordable in the case of lesser-used languages.

In spite of the fact that word alignment has been widely used for more than a decade within the NLP community to produce bilingual lexicons (Wu and Xia, 1994) and several experts claimed that such resources might also be useful for lexicographic purposes (e.g. Bertels et al., 2009), as far as we know, this technique has not been exploited in large-scale lexicographic projects, yet (e.g. Atkins and Rundell, 2008).

Our earlier experiments has shown that although word alignment has advantages over more traditional approaches, there are also some difficulties that have to be dealt with: Proto-dictionaries comprise incorrect translation candidates, as well, and the method in itself does not handle multi-word expressions. In fact, in a given parallel corpus the number of incorrect translation candidates strongly depends on the size of the proto-dictionary, as there is a trade-off between precision and recall.

Accordingly, our recent objective is to design and implement a dictionary query system that is apt to exploit the additional benefits of the method and overcome the disadvantages of it. According to our expectations such a system renders the proto-dictionaries helpful for not only lexicographers, but also for ordinary dictionary users.

The proto-dictionaries are available at: http://efnilex.efnil.org

2. Generating proto-dictionaries

2.1. *Input data*

Since the amount of available parallel data is crucial for this approach, in the first phase of the project we have experimented with two different language pairs. The Dutch-French language pair represents well-resourced languages while the Hungarian-Lithuanian language pair represents medium density languages. As for the former, we have exploited the French-Dutch parallel corpus, which forms subpart of the Dutch Parallel Corpus (Macken et al., 2007). It consists of 3,606,000 French tokens, 3,215,000 Dutch tokens and 186,945 translation units. The size of the parallel corpora is given in terms of translation units instead of in terms of sentence pairs, since many-to-many alignment among source and target sentences was allowed (TUs). As for Hungarian and Lithuanian we have built a parallel corpus comprising 4,189,000 Hungarian and 3,544,000 Lithuanian tokens and 262,423 TUs.

Because our original intention is to compile dictionaries covering the every-day language we have decided to focus on literature while collecting the texts. However, due to the scarce availability of parallel texts we made some concessions that might be questionable from a translation point of view. First, we did not confine ourselves purely to the literary domain: Philosophical works were also included. Secondly, instead of focusing on direct translations between Lithuanian and Hungarian we have relied mainly on translations from a third language. Thirdly, we have treated every parallel text alike, regardless of the direction of the translation, although the DPC contains that information.

2.2. *The generation process*

As already has been mentioned in the introduction, word alignment in itself deals only with one-token units. A detailed description of the generation process of such proto-dictionaries has been given in previous papers (e. g. Héja, 2010). In the present paper we confine ourselves to a schematic overview. In the first step the lemmatized versions of each input text have been created by means of morphological analysis and disambiguation. The analysis of the Lithuanian texts was performed by the Lithuanian Centre of Computational Linguistics (Zinkevičius et al., 2005). The Hungarian texts were annotated with the tool-chain of the Research Institute for Linguistics, HAS (Oravecz and Dienes, 2002).

In the second step parallel corpora have been created. We used Hunalign (Varga et al., 2005) for sentence alignment.

In the next step word alignment has been performed with GIZA++ (Och and Ney, 2003). During word alignment GIZA++ builds a dictionary-file that stores translation candidates, that is, source and target language lemmata along with their translation probabilities. Translational probability is the estimation of the conditional probability of the target word given the source word: p(L_t|L_s). We used this dictionary file as the starting point to create the proto-dictionaries.

In the fourth step the proto-dictionaries have been created. Only the most likely translation candidates were kept on the basis of some suitable heuristics, which has been developed while evaluating the results manually.

Finally, the relevant example sentences were provided in a concordance to give hints on the use of the translation candidates.

2.3. *Trade-off between precision and recall*

At this stage of the workflow some suitable heuristics need to be introduced to find the best translation candidates without the loss of too many correct pairs. Therefore, several evaluations were carried out.

It is important to note that throughout the manual evaluation we have focused on lexicographically useful translation candidates instead of perfect translations. The reason behind this is twofold. First, translation synonymy is rare in general language (e.g. Atkins and Rundell, 2008: 467), thus other semantic relations, such as hyponymy or hyperonymy, were also considered. Secondly, since the word alignment method does not handle MWEs in itself, partial matching between SL and TL translation candidates occurs frequently. In either case, provided example sentences make possible to find the right translation.

We considered three parameters when searching for the best translations: The translational probability, the source language lemma frequency and the target language lemma frequency (p_{tr}, F_s and F_t, respectively).

The lemma frequency had to be taken into account for at least two reasons.
(1) A minimal amount of data was necessary for the word alignment algorithm to be able to estimate the translational probability. (2) In the case of rarely used TL lemmas the alignment algorithm might assign high translational probabilities to incorrect lemma pairs if the source lemma occurs frequently in the corpus and both members of the lemma pair recurrently show up in aligned units.

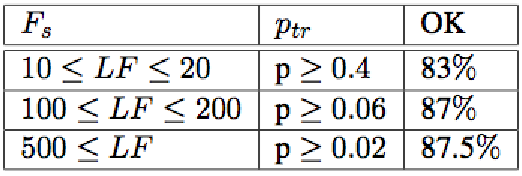
Results of the first evaluation showed that translation pairs with relatively low frequency and with a relatively high translational probability yielded cc. 85% lexicographically useful translation pairs. Although the precision was rather convincing, it has also turned out that the size of the resulting proto-dictionaries might be a serious bottleneck of the method (Héja, 2010). Whereas the targeted size of the dictionaries is between 15,000 and 25,000 entries, the proto-dictionaries comprised only 5,521 Hungarian-Lithuanian and 7,007 French-Dutch translation candidates with the predefined parameters. Accordingly, the coverage of the proto-dictionaries should be augmented.

According to our hypothesis in the case of more frequent source lemmata even lower values of translational probability might yield the same result in terms of precision as in the case of lower frequency source lemmata. Hence, different evaluation domains need to be determined as a function of source lemma frequency. That is: (1) The refinement of the parameters yields approximately the same proportion of correct translation candidates as the basic parameter setting, (2) The refinement of the parameters ensures a greater coverage.

Detailed evaluation of the French-Dutch translation candidates confirmed the first part of our hypothesis. We have chosen a parameter setting in accordance with (1) (see Table 1). 6934 French-Dutch translation candidates met the given conditions. 10 % of the relevant pairs were manually evaluated.

The results are presented in Table 1. '*OK*' denotes the lexicographically useful translation candidates. For instance, the first evaluation range (1st row of Table 1) comprised translation candidates where the source lemma occurs at least 10 times and at most 20 times in the parallel corpus. With these parameters only those pairs were considered where the translation probability was at least 0.4. As the 1st and 2nd rows of Table 1 show, using different p_{tr} values as cut-off parameters give similar results 87%), if the two source lemma frequencies also differ.

Table 1. Evaluation results of the refined French-Dutch proto-dictionary.

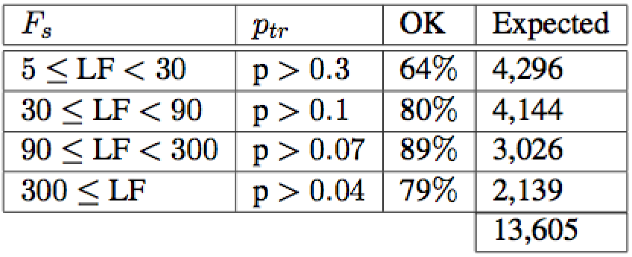


F_s	p_{tr}	OK
$10 \leq LF \leq 20$	$p \geq 0.4$	83%
$100 \leq LF \leq 200$	$p \geq 0.06$	87%
$500 \leq LF$	$p \geq 0.02$	87.5%

Manual evaluation of the Hungarian-Lithuanian translation candidates yielded the same result. We have used this proto-dictionary to confirm the 2nd part of our hypothesis, that is, the refinement of these parameters may increase the size of the proto-dictionary. Table 2 presents the results. *'Expected'* refers to the expected number of correct translation candidates, estimated on the basis of the evaluation sample. 800 translation candidates were evaluated altogether, 200 from each evaluation domain.

As Table 2 shows, it is possible to increase the size of the dictionary through refining the parameters: with fine-tuned parameters the estimated number of useful translation candidates was 13,605 instead of 5,521.

Table 2. Evaluation results of the refined Hungarian-Lithuanian proto-dictionary.



F_s	p_{tr}	OK	Expected
$5 \leq \mathrm{LF} < 30$	$p > 0.3$	64%	4,296
$30 \leq \mathrm{LF} < 90$	$p > 0.1$	80%	4,144
$90 \leq \mathrm{LF} < 300$	$p > 0.07$	89%	3,026
$300 \leq \mathrm{LF}$	$p > 0.04$	79%	2,139
			13,605

However, we should keep in mind when searching for the optimal values for these parameters that while we aim at including as many translation candidates as possible, we also expect the generated resource to be as clean as possible. That is, in the case of proto-dictionaries there is a trade-off between precision and recall: the size of the resulting proto-dictionaries can be increased only at the cost of more incorrect translation candidates.

This leads us to the question of what parameter settings are useful for what usage scenarios? We think that the proto-dictionaries generated by this method with various settings match well different user needs. For instance, when the settings are strict so that the minimal frequencies and probabilities are set high, the dictionary will contain less translation pairs, resulting in high precision and relatively low coverage, with only the most frequently used words and their most frequent translations. Such a dictionary is especially useful for a novice language learner.

Professional translators are able to judge whether a translation is correct or not. They might be rather interested in special uses of words, lexicographically useful but not perfect translation candidates, and more subtle cross-language semantic relations, while at the same time, looking at the concordance provided along with the translation pairs, they can easily catch wrong translations which are the side-effects of the method. This kind of work may be supported by a proto-dictionary with increased recall even at the cost of a lower precision.

Thus, the Dictionary Query System described in Section 5 in more detail, should be customizable: It should be able to support various user needs.

However, user satisfaction has to be evaluated in order to confirm our hypothesis. It forms part of our future tasks.

3. Including more subtle information in the proto-dictionaries

3.1. *The internal representation of the parallel corpus*

After confirming that the proposed method is able to generate proto-dictionaries with appropriate coverage and precision, we focused on the retrieval of more subtle data. For instance, it would be rather useful, if part-of-speech information were available, or the automatically attained translation pairs were assigned to typical text types in which they occur. Thus, we have converted the parallel corpus into XML-format, which contains all the relevant information in a structured way, which can be extracted in different ways when needed.

3.2. *Harmonizing the morphological annotation*

Since morphological analyzers vary from language to language, using different annotations, the morphological information has to be harmonized so that it can be processed in a uniform way later on, regardless of the previous processing steps.

We have kept basic part-of-speech information and the case information for Hungarian. Every other morphological annotation was omitted.

The resulting parallel corpus comprised 1,045,467 Hungarian tokens and 1,224,675 Slovenian tokens. It consisted of 38,791 TUs. In the next step a parallel corpus containing part-of-speech information was produced based on the parallel XML files. POS-tags might help disambiguating among different senses. For instance, the Hungarian lemma *bár* has multiple meanings. As a noun it is a type of pub whereas as a conjunction it means *but*. This sense distinction is clearly reflected by the Slovenian translations: As a noun it is translated as *bar* and *locale*, while as a conjunction it is translated as *čeprav* (’although’), *četudi* (’even if’) or *vendar* (’however’).

4. Semantic relations between source words and target words

According to our hypothesis translational probabilities and the ratio of the frequencies of the source and target lemmata provide useful hints on the semantic relations between the source and target lemmata. This supposition is in accordance with Dyvik (2002), whose objective was to build a WordNet based on parallel data.

They start out from the observation that ’Translations come about when translators evaluate the degree of interpretational equivalence between linguistic expressions in specific contexts. In many ways such evaluations, made without any theoretical concerns in mind, seem more reliable as sources of semantic information than the careful paraphrases of the semanticist or the meaning descriptions of the lexicographer.’ Accepting this observation we think that the basic assumptions behind their method can be easily interpreted in translational terms.

(1) ’Semantically closely related words ought to have strongly overlapping sets of

translations.'

In translational terms semantically closely related words are translational synonyms. In our framework this can be formulated in two ways:

(a) Translational probability is high and the frequencies of the source and target lemmata are close. (b) The straight and reverse translational probabilities are both high.

As Atkins and Rundell (2008:467) states, 'The perfect translation – where an SL word exactly matches a TL word – is rare in general language, except for the names of objects in the real world (natural kind terms, artefacts, places, etc.)'. Manual evaluation of Slovenian and Hungarian translation pairs (p_{tr}=1, frequency ratio is less than 3) yielded the result that out of 136 translation pairs 104 were noun-to-noun translations. Besides the semantic categories mentioned above proper names, illnesses, professions could obviously detected, but surprisingly, a couple of abstract nouns were also included, such as *ihlet* ('inspiration') and *botrány* ('scandal'). However, because translational synonymy is rare, other semantic relations have to be considered, too.

(2) 'Words with wide meanings ought to have a higher number of translations than words with narrow meanings.'

From a translation point of view it is rather important to know whether the meaning of the translation is more general or more restricted than that of the source word. In the latter case the context has to be paid great attention when selecting the right translation. Provided example sentences facilitate this choice. Moreover, hints can be given based on the source and target frequencies and the translational probabilities. If the source word is more frequent than the target word and the translational probability is relatively high then the meaning of the source word is wider than the meaning of the target word. Conversely, the higher frequency of the target word along with a great value of translational probability implies that the target word has a wider meaning. For instance, the Lithuanian word *puikus* (freq = 1003) is assigned several Hungarian translations with various frequencies, such as *jó* ('good' or 'great'), *kiváló* ('outstanding'), *pompás* ('gorgeous'), *kitűnő* ('excellent') etc. The frequency of the most likely translation *jó* (freq = 6871) implies that the target word's meaning is more general. Conversely, the lower frequencies (135, 187, 131, respectively) of the other translations imply that their meanings are more restricted.

(3) 'Furthermore, if a word *a* is a hyponym of a word *b* (such as *tasty* of *good*, for example), then the possible translations of *a* ought to be a subset of the possible translations of *b*.'

According to our hypothesis if the sum of the target lemma frequencies is close to the source lemma frequency and the sum of their translation probabilities is high then the target lemmata represent submeanings of the source word. The submeanings might be related or homonyms. In the previous case it might be said that the source lemma is the hyperonym of the target lemma. Unfortunately, at the present stage of research we cannot automatically tell apart the two cases from each other. For example, the Lithuanian lemma *kinas* is translated as *kínai* ('Chinese') and *mozi* ('cinema') at the same time. Another example is the Slovenian word *vrata* (freq = 542) the Hungarian translations of which are *ajtó* ('door', freq = 430), *kapu* ('gate', freq = 67), *bejárat* ('entrance', 39) the sum of their translation probability is 89%.

5. The dictionary query system

As earlier has been mentioned, the proposed method has several benefits compared to more traditional approaches: (1) A parallel corpus of appropriate size guarantees that the most relevant translations be included in the dictionary. (2) Based on translational probabilities it is possible to rank translation candidates ensuring that the most likely used translation variants go first within an entry. (3) All the relevant example sentences from the parallel corpora are easily accessible facilitating the selection of the most appropriate translations from possible translation candidates.

The Dictionary Query System presents some novel features to exploit the above advantages. On the one hand, users can select the best proto-dictionary for their purposes on the Cut Board Page. On the other hand, the innovative representation of the generated bilingual information helps to find the best translation for a specific user in the Dictionary Browser Window.

5.1. Customizable proto-dictionaries: the Cut Board Page

The dictionary can be customized on the Cut Board Page. Two different charts are displayed here showing the distribution of all word pairs of the selected proto-dictionary.

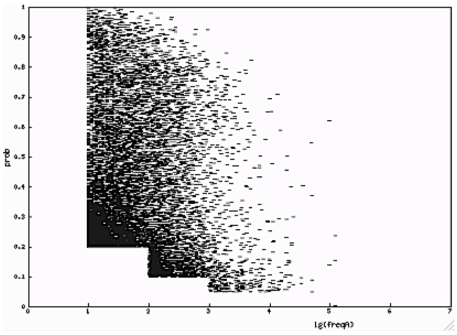


Figure 1. The customized dictionary: the distribution of the Lithuanian-Hungarian translation candidates. Logarithmic frequency of the source words on the *x*-axis, translation probability on the *y*-axis.

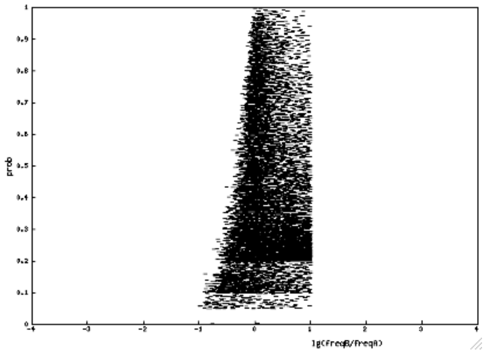


Figure 2. The customized dictionary: the distribution of the candidates. Logarithmic frequency ratio of the source and target words on the x-axis, translation probability on the y-axis.

Figure 1 visualizes the distribution of the logarithmic frequency of the source words and the relevant translation probability for each word pair, selected by the given custom criteria.

Figure 2 visualizes the distribution of the logarithmic frequency ratio of the target and source words and the corresponding translation probability for each word pair, selected by the given custom criteria.

Proto-dictionaries are customizable by the following criteria:

(1) Maximum and minimum ratio of the relative frequencies of the source and target words (left and right boundary on Plot 2 (Figure 2).

(2) Overall minimum frequency of the source or the target words (left boundary on Plot 2 (Figure 2).

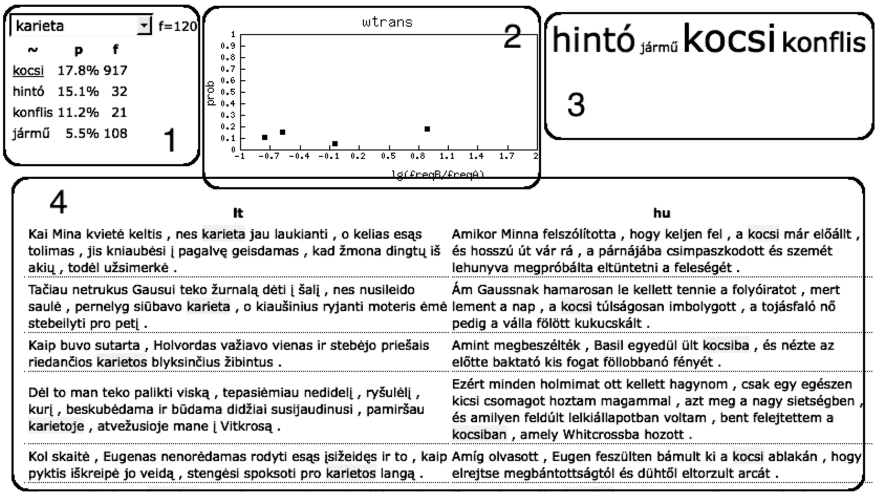
(3) Overall minimum translation probability (bottom boundary on both plots).

(4) Several more cut-off intervals can be defined in the space represented by Plot 1: Word pairs falling in rectangles given by their left, right and top boundaries are cut off.

After submitting the given parameters the charts are refreshed giving a feedback to the user and the parameters are stored for the session, i. e. the dictionary page shows only word pairs fitting the selected criteria.

5.2. *Dictionary Browser*

As Figure 3 illustrates, the Dictionary Browser displays four different types of information.



~	p	f
kocsi	17.8%	917
hintó	15.1%	32
konflis	11.2%	21
jármű	5.5%	108

lt	hu
Kai Mina kvietė keltis , nes karieta jau laukianti , o kelias esąs tolimas , jis kniaubėsi į pagalvę geisdamas , kad žmona dingtų iš akių , todėl užsimerkė .	Amikor Minna felszólította , hogy keljen fel , a kocsi már előállt , és hosszú út vár rá , a párnájába csimpaszkodott és szemét lehunyva megpróbálta eltüntetni a feleségét .
Tačiau netrukus Gausui teko žurnalą dėti į šalį , nes nusileido saulė , pernelyg siūbavo karieta , o kiaušinius ryjanti moteris ėmė stebeilyti pro petį .	Ám Gaussnak hamarosan le kellett tennie a folyóiratot , mert lement a nap , a kocsi túlságosan imbolygott , a tojásfaló nő pedig a válla fölött kukucskált .
Kaip buvo sutarta , Holvordas važiavo vienas ir stebėjo priešais riedančios karietos blyksinčius žibintus .	Amint megbeszélték , Basil egyedül ült kocsiba , és nézte az előtte baktató kis fogat föllobbanó fényét .
Dėl to man teko palikti viską , tepasiėmiau nedidelį , ryšulėlį , kurį , beskubėdama ir būdama didžiai susijaudinusi , pamiršau karietoje , atvežusioje mane į Vitkrosą .	Ezért minden holmimat ott kellett hagynom , csak egy egészen kicsi csomagot hoztam magammal , azt meg a nagy sietségben , és amilyen feldúlt lelkiállapotban voltam , bent felejtettem a kocsiban , amely Whitcrossba hozott .
Kol skaitė , Eugenas nenorėdamas rodyti esąs įsižeidęs ir to , kaip pyktis iškreipė jo veidą , stengėsi spoksoti pro karietos langą .	Amíg olvasott , Eugen feszülten bámult ki a kocsi ablakán , hogy elrejtse megbántottságtól és dühtől eltorzult arcát .

Figure 3. The Dictionary Browser.

(1) List of the translation candidates ranked by their translation probabilities. This guarantees that most often used translations come first in the list (from top to bottom). Absolute corpus frequencies are also displayed.

(2) A plot displaying the distribution of the possible translations of the source word according to translation probability and the ratio of corpus frequency between the source word and the corresponding translation candidate.

(3) Word cloud reflecting semantic relations between source and target lemmata. Words in the word cloud vary in two ways.

First, their *size* depends on their translation probabilities: the higher the probability of the target word, the bigger the font size is.

Secondly, *colours* are assigned to target words according to their frequency ratios relative to the source word: less frequent target words are cool-coloured (dark blue and light blue) while more frequent target words are warm-coloured (red, orange). Target words with a frequency close to that of the source word get gray colour.

(4) Provided example sentences with the source and target words highlighted, displayed by clicking one of the translation candidates.

Semantic relations are represented by colours. For instance, the Lithuanian lemma *karieta* has four Hungarian equivalents: *kocsi* (word with general meaning, e.g. ’car’, ’railway wagon’, ’horse-drown vehicle’), *hintó* (’carriage’), *konflis* (’a horse-drawn vehicle for public hire’), *jármű* (’vehicle’). The various colours of the candidates indicate different semantic relations: the red colour of *kocsi* marks that the meaning of the target word is more general than that of the source word. Conversely, the dark blue colour of *konflis* shows that the meaning of the target word is more special. However, this hypothesis should be tested in the future, which makes part of our future work.

6. Conclusions and future work

Previous experiments have proven that corpus-driven bilingual resources generated fully by automatic means are apt to facilitate lexicographic work when compiling bilingual dictionaries. We think that the proto-dictionaries generated by this technique with various settings match well different user needs, and consequently, besides lexicographers, they might also be useful for end users. A possible future work is to further evaluate the dictionaries in real world use cases.

Some new assumptions can be formulated which connect the statistical properties of the translation pairs, for example, their frequency ratios and the cross-language semantic relations between them. Based on the generated dictionaries such hypotheses may be further examined in the future.

In order to demonstrate the generated proto-dictionaries, we have designed and implemented an online dictionary query system, which exploits the advantages of the data-driven nature of the applied technique. It provides different visualizations of the possible translations. By presetting different selection criteria the contents of the dictionaries are customizable to suit various usage scenarios.
The dictionaries are publicly available at: http://efnilex.efnil.org.

Note

[1] The work presented here is part of the pilot project EFNILEX financed by EFNIL. Our thanks go to Judit Kuti and Piroska Lendvai, Justina Lukaseviciute and Beatrix Tölgyesi for their contribution to the manual evaluation.

References

Atkins, B. T. and M. Rundell 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Bertels, A., C. Fairon, J.Tiedemann and S. Verlinde 2009. ‘Corpus parallèles et corpus ciblés au secours du dictionnaire de traduction.’ *Cahiers de lexicologie*, 94:199–219.

Dyvik, H. 2002. ‘Translations as Semantic Mirrors: From Parallel Corpus to Wordnet 1.’ In *Section for Linguistic Studies Scientific Papers*, University of Bergen.

Erjavec, T., C. Ignat, B.Pouliquen and R. Steinberger 2005. ‘Massive Multi-Lingual Corpus Compilation: Acquis Communautaire and Totale.’ In Z. Vetulani (ed.), *Human language technologies as a challenge for computer science and linguistics: 2nd Language & Technology Conference, April, 21-23, 2005, Poznań, Poland: proceedings*. Poznań : Wydawnictwo Poznańskie, 32–36.

Héja, E. 2010. ‘The Role of Parallel Corpora in Bilingual Lexicography.’ In N. Calzolari, K. Choukri et al. (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10), Valletta, Malta, May*. ELRA, 2798–2805.

Macken, L., J. Trushkina, H. Paulussen, L. Rura, P. Desmet and W. Vandeweghe 2007. ‘Dutch Parallel Corpus: A Multilingual Annotated Corpus.’ In *On-line Proceedings of*

Corpus Linguistics 2007, 27-30 July 2007, Birmingham, United Kingdom.

Martin, W. 2007. ‘Government Policy and the Planning and Production of Bilingual Dictionaries : The Dutch Approach as a Case in Point.’ *International Journal of Lexicography* 20.3: 221–237.

Och, F. J. and H. Ney 2003. ‘A Systematic Comparison of Various Statistical Alignment Models.’ *Computational Linguistics* 29.1: 19–51.

Oravecz, C. and P.Dienes 2002. ‘Efficient Stochastic Part-of-Speech Tagging for Hungarian.’ In M. González Rodríguez and C. P. Suarez Araujo (eds.), *Third international conference on language resources and evaluation, 29th, 30th & 31st May 2002, Las Palmas de Gran Canaria, (Spain): proceedings*. Paris: European Language Resources Association, 710–717.

Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón and V. Nagy 2005. ‘Parallel Corpora for Medium Density Languages.’ In *Recent Advances in Natural Language Processing (RANLP 2005)*, 590–596.

Wu, D. and X. Xia 1994. ‘Learning an English-Chinese Lexicon from a Parallel Corpus.’ In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*. 206–213.

Zinkevičius, V., V. Daudaravičius and E. Rimkutė 2005. ‘The Morphologically Annotated Lithuanian Corpus.’ In M. Langemets and P. Penjam (eds.)*, The second Baltic Conference on Human Language Technologies : proceedings, April 4-5, 2005, Tallinn, Estonia.* Tallinn: Institute of Cybernetics, Tallinn University of Technology: Institute of the Estonian Language, 365–370.