Enikő Héja / Tamás Váradi

EFNILEX online dictionaries

1. Introduction

EFNILEX is a lexicographical project launched by EFNIL in 2008 to explore to what extent language technology methods are able to support the efficient and cost-effective creation of bilingual dictionaries. The targeted size of the dictionaries was between 15,000 and 25,000 entries covering everyday language vocabulary. The aim of the present paper is to provide an overview of the first three years of EFNILEX.[1]

The political relevance of the project is given by the fact that the European Union wishes to contribute to policies aimed at the preservation and strengthening of the multilingualism of Europe and the plurilingualism of its citizens. This goal implies that as many languages as possible should be:

(i) used in as many domains, functions and situations as possible;
(ii) involved in cross-border European and global communication and information exchange, e.g. through the internet;
(iii) learned and used by as many users as possible, both native and non-native speakers.

The above objectives imply that special attention has to be paid to lesser used language pairs, where – due to the low demand – even dictionaries of appropriate size and quality are hardly available.

In the first stage of the project current methods for the automatic creation of lexical resources were explored. This task is principally a survey of related research, focusing on two main directions:

(i) research aiming at the creation of a third dictionary by linking two existing bilingual lexical databases;
(ii) research on the automatic construction of bilingual resources from aligned parallel corpora.

In either case, there are no methods that could enable the wholly automatic production of dictionaries according to the state of the art. Thus, the production of a proper lexicographical resource necessarily requires a post-editing phase. Hence, the objective of the project was to provide lexicographers with resources diminishing as much as possible the amount of labour required to prepare full-fledged dictionaries, not excluding the possibility that such resources might be useful for end-users, too. These kind of resources will be referred to as proto-dictionaries henceforward. The selection of the most promising method will be described in Section 2.

[1] The present paper is a summarization of the work described in Héja (2010ab) and Héja/Takács (to appear 2012).

After choosing the method that best suits our preferences, proof-of-concept experiments were performed to confirm the viability of the proposed approach and to explore the related difficulties. Throughout the experiments we considered both lesser-used and well-resourced language pairs, such as Hungarian and Lithuanian and Dutch and French, respectively. The workflow is described in Section 3.

In Section 4 an overview is given of the pros and cons of the proposed approach. The applied technique proved to be promising for several reasons. Most importantly, if input data of appropriate size are available, the proposed approach is able to guarantee that the most relevant translations are included in the dictionary.

Moreover, it is possible to rank translation candidates based on automatically attained translational probabilities, which ensures that the most likely translation variants go first within an entry. A further advantage is that all the relevant example sentences from the parallel corpora are easily accessible, thus facilitating the selection of the most appropriate translations. However, we also had to face some difficulties. On the one hand, the proposed method is not able to handle multi-word expressions in itself. On the other hand, although the results of the proof-of-concept experiments turned out to be rather good in terms of precision, the coverage of the resulting dictionaries should be increased further.

Accordingly, in the third stage of the project we have tried to come up with a solution to the problem of low coverage. Two alternatives are described in Section 5.

Finally, a dictionary query system was designed and implemented that is able to compensate for the drawbacks of the selected method and to extend its advantages even further. Although no user case study has been performed, according to our expectations a proper query system is able to render the automatically generated resources useful for not only lexicographers but for end-users, too. The automatically generated online dictionaries are available at http://efnilex.efnil.org. See Section 6 for more details on the Dictionary Query System.

2. Selecting the method

2.1 Expectations toward the EFNILEX bilingual dictionaries

The selection of the most appropriate technique was guided by two interrelated aspects: the ability to *reduce the amount of human labour* needed to compile dictionaries and the suitability to *decrease* or *eliminate the reliance on human intuition* during lexicographic work.

Whereas the first objective is rather straightforward, that is, the proposed method should be as automatic as possible, the second objective needs to be explained in more detail. However, a few definitions are in order before giving an overview of the possible negative effects of relying on human intuition during lexicographic work.

The term *lemma* refers to the stem of a wordform, regardless of its meaning. For instance, *nail* is a lemma no matter if it denotes the bodypart or the thin pointed peace of metal. As opposed to lemma, the term ‘linguistic unit’ (LU) denotes a lemma-meaning pair. Accordingly, the lemma *nail* is associated with at least two LUs based on the different meanings given above.

Following Atkins/Rundell (2008) the dictionary building process can be decomposed into three distinct phases (see Figure 1).

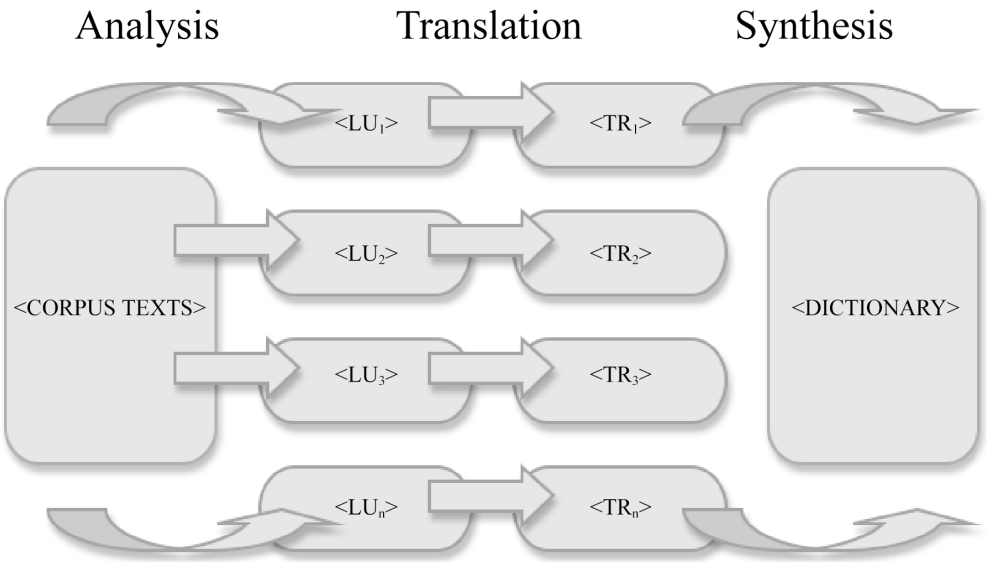

Figure 1: A schematic overview of the dictionary building process

In the *analysis stage* a relevant headword list of the source language has to be created. This stage includes both the selection of the relevant lemmata and the characterization of word senses belonging to each of the lemmata (LUs). Hence, an inherent part of this stage is making decisions on which alternative senses are to be included in the source language side of the dictionary. The exploitation of existing monolingual dictionaries, wordnets or monolingual corpora might facilitate the compilation of such a headword list the entries of which can be translated into the TL in the next stage.

Since the goal of dictionary use is to find the best translations for the given contexts, it is important that the alternative senses of an SL lexeme could be assigned with high agreement to words in context even in the source language. Unfortunately, finding the relevant meanings of words in contexts is not at all obvious. This claim is confirmed by at least two experiments (e.g. Véronis 2003; Kuti et al. 2010) where human annotators were asked to tag words in contexts with their relevant meaning on the basis of monolingual sense inventories, such as Petit Larousse explanatory dictionary**,** Hungarian WordNet (Miháltz et al 2008) or the Hungarian Explanatory Dictionary. All these experiments resulted in low values of inter-annotator agreement. These results clearly imply that these sense inventories are not suitable for the sense-tagging of tokens in their contexts. That is, such databases cannot be trustworthily used for finding the best translations in contexts.

Certainly, the experiments above are not capable of proving that handcrafted sense inventories (e.g. those built in the framework of the CLVV-projects, Martin 2007) are not suited for obtaining high inter-annotator agreement. However, the results underpin that distributional data have to be carefully explored and taken into consideration when constructing such databases. Since building sense inventories manually that exploit linguistic information as much as possible is rather expensive, this approach is typically not affordable in the case of lesser-used languages.

During the *transfer stage* the linguistic units making up the headword list are translated into the target language. According to Atkins/Rundell (2008, 135) although "the relationship of synonymy should ideally hold between the headword and its target-language equivalent" applying synonymy as a criterion is impossible in most cases, as "it is difficult to find convincing examples of synonyms, because true synonyms are extremely rare, if they exist at all. The nearest you get is usually a pseudo-synonym." A more viable approach to translational equivalency is to hunt for direct translations, i.e. for translations "that suit most of the contexts" (Atkins/Rundell 2008, 464) of the source language expression. Hence, in most cases contexts (at least) of the SL expression have to be thoroughly explored to be able to determine the best translation. However, exploiting an appropriately characterized monolingual sense inventory yields only a partial solution to the problem of how to find the best translation for a given source language expression, since lexicographers still have to make use of their intuition when selecting the ideal translation out of the possible translation candidates.

2.2 Word alignment on parallel corpora

According to our expectations applying automatic word alignment on parallel corpora might overcome the difficulties listed above.

First, it is cost-effective: instead of expoiting refined handcrafted monolingual resources, this method uses parallel texts as starting point.

Secondly, it helps to diminish the reliance on human intuition during lexicographic work. Accordingly, in this approach, neither source language nor target language LUs are extracted directly by lexicographers from the corpus. Instead, LUs are determined by their contexts both in the SL and in the TL corpus and their translational equivalents provided by the parallel sentences. Furthermore, the corpus-driven nature of this method ensures that human insight is eliminated also when hunting for possible translation candidates, that is, when establishing possible pairings of the source language and the target language expressions. Moreover, the method ranks the translation candidates according to how likely they are, based on automatically attained translational probabilities. This in turn renders it possible to determine which sense of a given lemma is the most frequently used, provided that distinct translations are available. Thus, representative corpora guarantee not only that the most important source lemmas will be included in the dictionary – as in traditional corpus-based lexicography – but also the translations of their most relevant senses.

The third great advantage of the proposed technique is that all the relevant natural contexts can be provided both for the source and for the target language. The contexts of the source language and the target language words could be exploited for multiple purposes.

First, they can be of great help in determining which translation variants should be used, thus enabling lexicographers to find the most appropriate translation on the one hand, and to describe the use of the target language expression in grammatical or collocational terms, on the other. Hence, the great amount of easily accessible natural contexts facilitates the creation of encoding or active dictionaries.

Secondly, different sub-senses of a headword can be characterized manually based on the retrieved contexts. Accordingly, dictionaries relying on such information can provide positive evidence for the user that all of these sub-senses are translated with the same lemma into the target language.

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/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/Rundell 2008).

3. Workflow

In the second phase of the project we have experimented with two lesser-used language pairs (Slovenian/Hungarian and Lithuanian/Hungarian) and with a language pair representing well-resourced languages (Dutch and French). The workflow comprised three main stages:

First, resources and language-specific tools had to be collected to create the parallel corpora (see section 3.1).

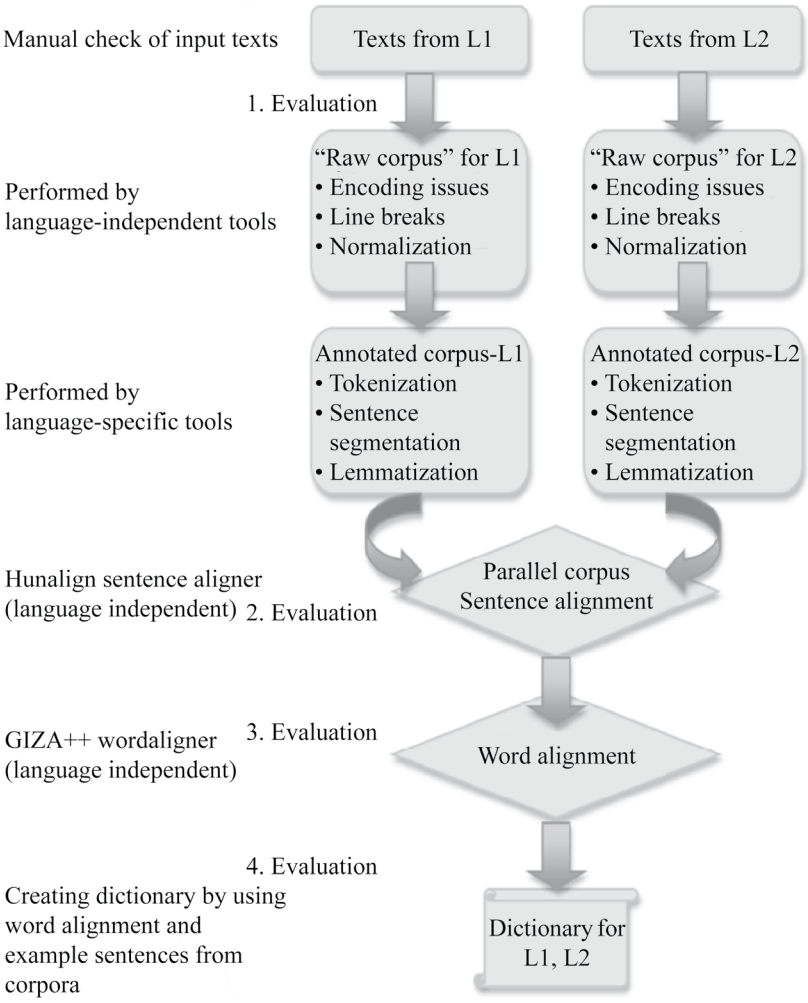

Figure 2: Workflow

Secondly, word alignment was carried out to generate the proto-dictionaries. Based on the preliminary manual evaluation of the Hungarian-Slovenian proto-dictionary some thresholds were set for some parameters based on which the unlikely translation candidates were filtered out. The same values were also applied in the case of Hungarian and Lithuanian (see section 3.2).

Finally, a more precise evaluation of the Hungarian-Lithuanian proto-dictionary was carried out manually by bilingual speakers, based on criteria that were also defined in this phase (see 3.3). Figure 2 presents a more detailed description of the work that has been accomplished. However, in the present paper only the phases listed above will be discussed in more detail.

3.1 Creation of parallel corpora

3.1.1 Collection of texts and tools

Since the objective of the project was to create dictionaries for everyday language vocabulary, we decided to focus on the genre fiction while collecting texts for our corpora. One of the main difficulties the project had to face was the scarce availability of general-domain parallel texts. As collecting direct translations yielded only a moderate success[2] we decided to gather texts translated from a third language.

Although national digital archives such as the Digital Academy of Literature[3] and the Hungarian Electronic Library[4] do exist in Hungary providing us with a wealth of electronically available texts, similar resources have not been found, neither for Slovenian nor for Lithuanian. Finally, we obtained sentence segmented and morphologically disambiguated texts from the Lithuanian Centre of Computational Linguistics, Vytautas Magnus University creator of the Lithuanian National Corpus (Rimkutė et al. 2007) and of the Lithuanian-English parallel corpus (Rimkutė et al. 2008).

Basic text-processing tasks were accomplished by means of language-specific tools accessible for all these three languages. As for Lithuanian, the analysis was carried out by the Lithuanian Centre of Computational Linguistics (Vytautas Magnus University). Slovenian texts were processed with the tool-chain available at the site of Jožef Stefan Institute[5] (Erjavec et al. 2005). Hungarian annotation was provided by the Part-Of-Speech Tagger of the Research Institute for Linguistics, HAS (Oravecz/Dienes 2002). As for Dutch and French we used the morphologically annotated and disambiguated Dutch-French parallel sub-corpus of the Dutch Parallel Corpus (DPC) created by the TLT-Centrale (Macken et al. 2007).

---

[2] For Lithuanian and Hungarian we did not find a significant amount of direct translations available in electronic form. In the case of Slovenian and Hungarian, we managed to gather a ca. 750,000-token corpus for each language through contacting several translators, publishers and the Slovenian Television.

[3] http://www.pim.hu/.

[4] http://mek.oszk.hu/.

[5] http://nl.ijs.si/jos/analyse.

3.1.2 Creation of parallel corpora

Sentence alignment was performed with hunalign (Varga et al. 2005). The lemmatized versions of the original texts served as input to sentence alignment to eliminate the problem of data sparseness resulting from rich morphology as much as possible.

Figure 3 shows the corpus size for each of the language pairs. The second column uses translational units (TUs) as a measure of corpus size instead of sentences. This is due to the fact that translations in parallel texts might merge or split up source language sentences, thus recognizing only one-to-one sentence mappings often entails loss of corpus data. Hunalign is able to overcome this difficulty by creating one-to-many or many-to-one alignments (i.e. 1:2, 1:3, 2:1, 3:1) between sentences.

LITHUANIAN-HUNGARIAN PARALLEL CORPUS		
LITHUANIAN	1,765,000 tokens	147,158 TUs
HUNGARIAN	2,121,000 tokens	147,158 TUs

SLOVENIAN-HUNGARIAN PARALLEL CORPUS		
SLOVENIAN	733,000 tokens	38,574 TUs
HUNGARIAN	666,000 tokens	38,574 TUs

Figure 3: Size of the parallel corpora in the case of lesser-used languages

The French-Dutch sub-corpus of the Dutch Parallel Corpus comprised 3,605,791 French tokens and 3,214,756 Dutch tokens. It consisted of 186,945 translational units.

3.2 Proto-dictionaries

This section presents how the list of translation candidates was generated (3.2.1), and how the most likely translation candidates were selected to produce the proto-dictionaries (3.2.2).

3.2.1 Creation of proto-dictionaries

The creation of the proto-dictionaries follows two main steps. The first step is word alignment for which the freely available tool GIZA++ (Och/Ney 2003) was used. To perform word alignment GIZA++ assigns translational probabilities to SL and TL lemma pairs. The translational probability is an estimation of the conditional probability of the target word given the source word, $P(W_{target}|W_{source})$ by means of the EM algorithm (Dempster et al. 1977). The retrieved lemma pairs with their translational probabilities served as the starting point for the proto-dictionaries. However, as the assigned translational probability strongly varies, at this stage we have many incorrect translation candidates. Therefore, some constraints had to be introduced to find the best translation candidates without the loss of too many correct pairs.

For this purpose, we focused on three parameters: the *translational probability*, the *source language lemma frequency* and the *target language lemma frequency*. First, the evalua-

tion of the Hungarian-Slovenian proto-dictionary was carried out. Due to the scarce availability of bilingual speakers for both Lithuanian and Slovenian, the first evaluation round provided the occasion for roughly estimating the settings of the above parameters. Then these parameters were applied to generate the Hungarian-Lithuanian and the French-Dutch proto-dictionaries and a more detailed evaluation was performed on them.

The lemma frequency had to be taken into account for at least two reasons. On the one hand, a minimal amount of data was necessary for the word alignment algorithm to be able to estimate the translational probability. On the other hand, 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.

3.2.2 Setting the parameters

The evaluation of a sample Hungarian-Slovenian proto-dictionary (5749 lemma pairs) has yielded the following findings:

1. Source language and target language members of lemma pairs should occur at least 5 times in order to have reliable amount of data when estimating probabilities.
2. If the translational probability is less than 0.5, the proportion of correct translation pairs drops considerably.

65% of the translation candidates with the corresponding parameters were correct translations. In the case of Hungarian-Lithuanian we also excluded translation candidates where either the Lithuanian or the Hungarian lemma occurred more than 100 times than the other in the whole parallel corpus.

Figure 4 indicates the number of translation candidates corresponding the to the given parameters. The second column of the table shows the number of expected correct translations, assuming that 65% of the translation candidates correct.

	NUMBER OF TRANSLATION-CANDIDATES ABOVE THE THRESHOLD	EXPECTED NUMBER OF CORRECT TRANSLATION-CANDIDATES
HUNGARIAN-SLOVENIAN	4969	3230
HUNGARIAN-LITHUANIAN	4025	2616

Figure 4: Expected size of the proto-dictionaries

Based on these parameters a detailed manual evaluation of the Hungarian-Lithuanian proto-dictionary was performed.

3.3 Detailed evaluation of the Hungarian-Lithuanian proto-dictionary

The evaluation was performed manually by bilingual speakers. Contrary to the usual evaluation methods, our basic objective was not to tell apart good translations from bad ones, instead, in accordance with our original purpose, we aimed at distinguishing between lexicographically useful and lexicographically useless translation candidates. The eligibility of this classification is clearly verified by the fact that there are completely correct translation pairs that are absolutely of no use for dictionary building purposes (e.g. specific proper names). On the other hand, incorrect translation pairs – in the strict sense – can be of great help for lexicographers, for example in the case of multiword expressions where the contexts provide lexicographers with sufficient amount of information to find the right translational equivalents.

In what follows, we will describe the categories used throughout the evaluation (3.3.1), then the methodology of the evaluation and the results will be presented (3.3.2).

3.3.1 Categories

The evaluation was based on two main categories: useful and useless translation candidates. Useful translation candidates comprised two subclasses.

(1) In the case of completely correct translation pairs no post-editing is needed.[6]

Example 1:

HUN: **gyümölcs** LIT: **vaisius** (fruit)

(2a) As opposed to completely correct translations, in the case of partially correct translations, post-editing has to be carried out, primarily due to incorrect lemmatization or partial matches in the case of multiword expressions. Example 2 illustrates the partial match in the case of a compound.

Example 2 (compounds):

HUN: **főfelügyelő** LIT: vyriausiasis **inspektorius** (chief inspector)

(2b) Example 3 gives an instance of partial match due to collocations.

Example 3 (collocations):

HUN: **bíborosi** testület LIT: Kardinolų **kolegiją** (cardinal college)

(2c) Partially correct translations might also result from slightly loose translations where no strong synonymy holds between the translation candidates. However, taking into consideration that synonymy in the strict sense is quite rare across languages, members of this class might yield quite useful clues on SL and TL lemmas with related meanings, which can, nevertheless, be substituted in certain contexts. Example 4 illustrates the semantic relation of hyperonymy.

Example 4:

HUN: **lúdtoll** (literally: goose-feather) LIT: **plunksna** (literally: feather, pen) (intended meaning in both cases: quill pen)

[6] Translation candidates are boldfaced in the examples.

3.3.2 Evaluation methodology and the results

Out of the 4025 translation candidates with the parameters determined above 863 pairs were manually evaluated. Throughout the evaluation three intervals were distinguished based on the value of the translation candidates' translational probability. The translational probability of 520 candidates was within the range [0.5, 0.7) and 280 candidates' translational probability lied within [0.7, 1). The proportion of the number of translation candidates within these intervals reflects their actual proportion in our proto-dictionary. All the translation candidates with translational probability 1 (63 pairs) were also included in the evaluation. Figure 5 indicates the result of the evaluation.

	Useful candidates		**Useless candidates**	
P(tr)	OK	Post-editing	Irrelevant	Incorrect
[0.5, 0.7)	52.1%	32.9%	2.3%	12.7%
Sum	Σ 85%		Σ 15%	
[0.7, 1)	65.3%	31.9%	0.6%	2.2%
Sum	Σ 97.2%		Σ 2.8%	
1	38.0%	13.0%	49.0%	0%
Sum	Σ 51%		Σ 49%	

Figure 5: Results of the Hungarian-Lithuanian proto-dictionary

If we consider the sum of completely correct pairs and lexicographically useful candidates, we can state that 85% of the translation pairs is useful in the probability range between 0.5 and 0.7. This value goes up to 97.2% in the range between 0.7 and 1. Interestingly, translation pairs with the highest probability (1) are only 51% useful, and only 38% correct. This is due to the high proportion of not relevant proper names in this probability range.

4. Pros and cons

4.1 Pros: the treatment of multiple meanings

As it was pointed out earlier in subsection 2.2, one of the main benefits of the proposed method is that it enables the extraction of all the relevant translations available in the corpora, thus diminishing the role of human intuition during lexicographic process. Furthermore, it ranks the extracted translation candidates on the basis of their translational probabilities. These features imply that the proposed technique copes with related meanings more efficiently than traditional lexicography or lexicography based on monolingual corpora. The example of the Lithuanian lemma *aiškiai* illustrates that the provided parallel data could be of great help for lexicographers in describing the relevant conditions under which a target language expression could occur:

aiškiai	***tisztán***	[literally*: pure+ly*]	**(*clearly*)**
PERCEPTION	*lát, látszik, hall*	(*'see', 'seem', 'hear'*)	

aiškiai	***világosan***	[literally: *clear+ly*]	**(*clearly*)**
PERCEPTION	*lát, látszik, hall*	(*'see', 'seem', 'hear'*)	
COGNITION	*megért, gondolkodik*	(*'understand', 'think'*)	
COMMUNICATION	*beszél, válaszol*	(*'speak', 'answer'*)	
aiškiai	***láthatóan***	[literally: *visible+ly*]	**(*visibly*)**
EMOTION	*aggódik, mulattat, élvez, nem tetszik*	(*'be worried', 'amuse', 'enjoy', 'do not like'*)	
aiškiai	***jól***		**(*well*)**
PERCEPTION	*lát, látszik, hall*	(*'see', 'seem', 'hear'*)	

Although due to its size our corpus is not well suited for providing sufficient data for the complete description of these 33s, on the basis of the contexts several conclusions can be drawn. First, *tisztán*, *világosan* and *jól* can modify verbs of PERCEPTION. *Láthatóan* is clearly distinguishable, as it usually refers to the fact that the emotional change a person underwent was overt. *Világosan* is also commonly used with verbs of COGNITION and verbs of COMMUNICATION with the same meaning, i.e. the content of the communication is clearly comprehensible. As opposed to this, with verbs of COMMUNICATION *tisztán* would mean that the speech conveying the message was clearly pronounced. This kind of information can be of great help for a Lithuanian speaker who wants to make utterances in Hungarian.

4.2 Cons: coverage

As was previously mentioned, one of the main disadvantages of word alignment is that is does not handle multi-word expressions in itself. However, in the present paper we confine ourselves to the discussion of the other serious problem: low coverage of the proto-dictionaries.

Based on the evaluation of the sample, we might expect that 3549 translation candidates out of 4025 should be useful, which implies that the coverage should be augmented.

Two possible alternatives were investigated to overcome this difficulty: improving the size of the parallel corpora and the refinement of the parameters.

5. Enlargement of the proto-dictionaries

The size of the Hungarian-Lithuanian parallel corpus was doubled (262,423 TUs, 3,544,000 Lithuanian tokens, 4,189,000 Hungarian tokens). Unfortunately, the nearly double-sized parallel corpus yielded only 37% more translation candidates with the original parameters.

An alternative approach might be fine-tuning the parameters when fishing for useful translation candidates (*translational probability*, the *source lemma frequency* and the *target lemma frequency*). According to our hypothesis in the case of more frequent source lemmata even lower values of translational probability might yield the same re-

sult 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, and evaluation of the Hungarian-Lithuanian translation candidates confirmed the second part of our hypothesis. With refined parameters the estimated number of useful translation candidates amounts to 13,600 translation pairs instead of ca. 5,500 in the case of the Hungarian-Lithuanian proto-dictionary.

Unfortunately, the size of the proto-dictionaries can be increased only at the cost of more incorrect translation candidates. This leads us to the question: what parameter settings are useful for what usage scenarios? We think that the proto-dictionaries 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 fewer translation pairs including 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 should be customizable. However, user satisfaction has to be evaluated in order to confirm our hypothesis. It forms part of our future tasks.

6. The Dictionary Query System[7]

As has been mentioned earlier, 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.

[7] The features of the dictionary query system were invented in close cooperation with Dávid Takács.

6.1 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.

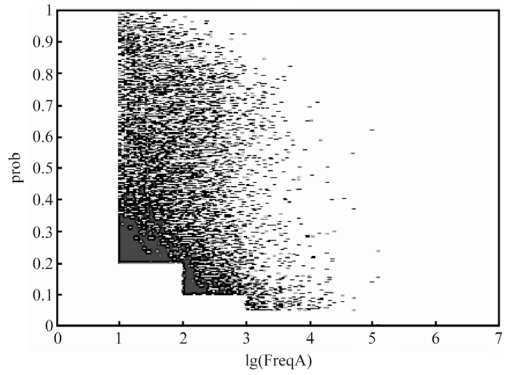

Figure 6: The customized dictionary: the distribution of the Lithuanian-Hungarian translation candidates. Logarithmic frequency of the source words on the *x*-axis, translation probability on the *y*-axis

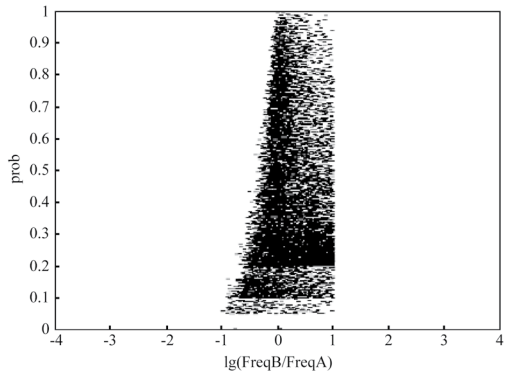

Figure 7: The customized dictionary: the distribution of the candidates. Logarithmic frequency ratio of the source and target words on the x-axis, translation probability on the y-axis

Figure 6 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 7 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 in Figure 7).
2. Overall minimum frequency of the source or the target words (left boundary in Figure 6).
3. Overall minimum translation probability (bottom boundary on both plots).
4. Several more cut-off intervals can be defined in the space represented by the first plot: 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.

6.2 The Dictionary Browser

As Figure 8 illustrates, the Dictionary Browser displays four different types of information.

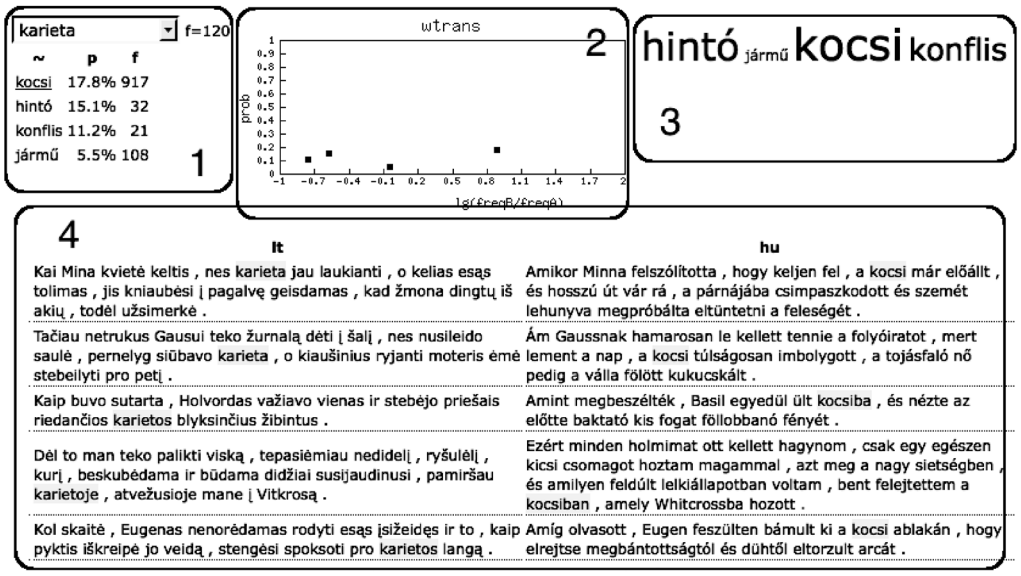

Figure 8: The Dictionary Browser

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. 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 with 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.

7. 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.

8. Acknowledgements

We are particularly grateful to František Čermák, John Simpson, Johan Van Hoorde, Jolanta Zabartskaite and Annemieke Hoorntje project partners for their contribution and valuable remarks.

Our thanks also go to Judit Kuti and Piroska Lendvai, Justina Lukaseviciute, Iván Mittelholcz, Bence Sárossy and Beatrix Tölgyesi for their contribution to the manual evaluation and to the collection of the texts. The Dictionary Query System was implemented by Dávid Takács.

References

Atkins, B.T.S./Rundell, M. (2008): *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

Bertels, A./Fairon, C./Tiedemann, J./Verlinde, S. (2009): Corpus parallèles et corpus ciblés au secours du dictionnaire de traduction. In: *Cahiers de Lexicologie* 94, 199-219.

Digitális Irodalmi Akadémia [Digital Academy of Literature]. Internet: www.pim.hu/.

Dempster, A.P./Laird, N.M./Rubin, D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm. In: *Journal of the Royal Statistical Society, Series B* 39 (1), 1-22.

Erjavec, T./Ignat, C./Pouliquen, B./Steinberger, R. (2005): Massive multi-lingual corpus compilation: Acquis Communautaire and totale. In: *Proceedings of the 2nd Language Technology Conference, April 21-23, 2005, Poznan, Poland*. Poznan, 32-36.

Héja, E. (2010a): Dictionary building based on parallel corpora and word alignment. In: Dykstra, A./Schoonheim, T. (eds): *Proceedings of the XIV. EURALEX International Congress, Leeuwarden, 6-10 July 2010*. Ljouwert: Fryske Akademy, 341-352.

Héja, E. (2010b): The role of parallel corpora in bilingual lexicography. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (*LREC '10*)*, 19-21 May 2010, Valletta, Malta*. Valletta, 2798-2805.

Héja, E./Takács, D. (to appear 2012): Automatically generated customizable online dictionaries. Systeme demonstration. In: *Proceedings of the EACL 2012, France. April 23-27*.

Kuti, J./Héja, E./Sass, B. (2010): Sense disambiguation – "Ambiguous sensation"? Evaluating sense inventories for verbal WSD in Hungarian. In: *Proceedings of LREC 2010 Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*. 23-29.

Magyar Elektronikus Könyvtár [Hungarian Electronic Library]: http://mek.oszk.hu/.

Martin, W. (2007): Government policy and the planning and production of bilingual dictionaries: The 'Dutch' approach as a case in point. In: *International Journal of Lexicography*, September 1, 20 (3), 221-237.

Miháltz, M./Hatvani, C./Kuti, J./Szarvas, G./Csirik, J./Prószéky, G./Váradi, T. (2008): Methods and results of the Hungarian WordNet Project. In: Tanács, A./Csendes, D./Vincze, V./Fellbaum, C./Vossen, P. (eds.): *Proceedings of the IVth Global WordNet Conference*. Szeged, 311-321.

Och, F.J./Ney, H. (2003): A systematic comparison of various statistical alignment models. In: *Computational Linguistics* 29 (1), 19-51.

Oravecz, C./Dienes, P. (2002): Efficient stochastic part-of-speech tagging for Hungarian. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas, 710-717.

Rimkutė, E./Daudaravičius, V./Utka, A./Kovalevskaitė, J. (2008): Bilingual parallel corpora for English, Czech and Lithuanian. In: *The Third Baltic Conference on Human Language Technologies 2007 Conference Proceedings*. Kaunas, 319-326.

Rimkutė, E./Daudaravičius, V./Utka, A. (2007): Morphological annotation of the Lithuanian corpus. In: *45th Annual Meeting of the Association for Computational Linguistics. Workshop Balto-Slavonic Natural Language Processing 2007 Conference Proceedings*. Praga, 94-99.

Varga, D./Németh, L./Halácsy, P./Kornai, A./Trón, V./Nagy, V. (2005): Parallel corpora for medium density languages. In: *Proceedings of the RANLP 2005*. Borovets, 590-596.

Véronis, J. (2003): Sense tagging: does it make sense? In: Wilson, A./Rayson, P./McEnery, T. (eds.): *Corpus linguistics by the lune: a festschrift for Geoffrey Leech*. Frankfurt a.M.: Peter Lang.

Wu, D. (1994): Learning an English-Chinese lexicon from a parallel corpus. In: *Proceedings of AMTA '94*. Columbia, MD, 206-213.