The Role of Parallel Corpora in Bilingual Lexicography

Enikő Héja
Dept. of Language Technology, Research Institute for Linguistics, HAS
P.O. Box. 360 H-1394 Budapest
eheja@nytud.hu

Abstract

This paper describes an approach based on word alignment on parallel corpora, which aims at facilitating the lexicographic work of dictionary building. Although this method has been widely used in the MT community for at least 16 years, as far as we know, it has not been applied to facilitate the creation of bilingual dictionaries for human use. The proposed corpus-driven technique, in particular the exploitation of parallel corpora, proved to be helpful in the creation of such dictionaries for several reasons. Most importantly, a parallel corpus of appropriate size guarantees that the most relevant translations are included in the dictionary. Moreover, based on the translational probabilities it is possible to rank translation candidates, which ensures that the most frequently used 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 from possible translation candidates. Due to these properties the method is particularly apt to enable the production of active or encoding dictionaries.

1. Introduction

This paper will investigate how language technology methods, in particular the exploitation of parallel corpora can contribute to dictionary building process, to render it as automatic as possible. This need shows up particularly in the case of medium-density language pairs, where – due to the low demand – investing in the production of dictionaries does not pay off for publishers.

The described work aims to produce medium-sized dictionaries covering everyday language vocabulary for Lithuanian and Hungarian. Slovenian and Hungarian was used as a test language pair.

According to the state of the art there are no methods that could enable the wholly automatic production of dictionaries. Thus, the production of a completely clean lexicographical resource with an appropriate coverage requires a post-editing phase. Hence, our goal is to provide lexicographers with resources diminishing the amount of labour required to prepare full-fledged dictionaries for human use as much as possible. These automatically generated resources will be referred to as core dictionaries henceforth.

The method we propose is based on statistical word alignment on sentence aligned parallel corpora. Although this approach has been widely used by the machine translation community for at least since 16 years (e.g. Wu, 1994) to improve the quality of dictionaries for machine translation purposes, as far as we know, parallel corpora and word alignment have not been exploited in lexicographical projects until now. Though this statement is difficult to verify, it is also confirmed by Atkins & Rundell (2008: 477): *"An appeal in January 2007 on the EURALEX discussion list for information about any dictionary publisher using a bilingual corpus in the editing of a bilingual dictionary produced no affirmative responses, but several working lexicographers commented on how useful such corpora could be"*.

The next section shortly presents the role of intuition in the traditional or corpus-based lexicography and the advantages of relying on sentence-aligned corpora while preparing dictionaries. The third section provides a brief description on the workflow: the creation of the parallel corpora and of the core dictionaries. It also describes the evaluation method and presents the results of the Hungarian-Lithuanian core dictionary. The fourth section illustrates how the proposed approach copes with related meanings. The last section summarizes the conclusions and some further tasks [5].

2. Advantages of Parallel Corpora in Dictionary Creation

2.1 The Task

The task of writing a bilingual dictionary might be conceived of as assigning the relevant language units of the target language (TL) to the relevant language units of the source language (SL). These language units ideally can be characterized as form-meaning pairs, and are usually referred to as *lexical units*. According to Atkins & Rundell (2008: 162-163) *„A headword in one of its senses is a lexical unit (or LU) [...]. LUs are the core building blocks of dictionary entries."* Thus the dictionary building process includes the characterization of the LUs to be included in the dictionary, and the selection of the most appropriate pairings between the source language and target language LUs. In some cases source language and target language LUs are described fully independently (e.g. CLVV project, Martin, 2007), in other cases only the source language LU list is built and the target language equivalents are produced by the means of translation afterwards. In either case the relation of *translational equivalence* has to hold between the corresponding entries. However, finding the ideal translations is not at all obvious, as Atkins & Rundell (2008: 467) asserts: *„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.)".
Moreover, in the case of *encoding dictionaries* (i.e. dictionaries providing speakers of the SL with information on how to express themselves in a foreign language) relevant contextual information of the TL also has to be included in the dictionary to give hints to users on how a TL expression should be used correctly.

2.2 The Dictionary Building Process

According to Atkins & Rundell (2008) the process of building a bilingual dictionary is threefold. At first, a relevant headword list of the source language has to be compiled. 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. In the latter case the production of a headword list is referred to as the *analysis stage*.
During the *transfer stage* the linguistic units making up the headword list are translated into the TL. However, it is important to keep in mind, especially during the creation of an encoding dictionary, that the translated LUs will be used in discourse. Thus, the safest translation has to be obtained, and possibly ranked at the first place in the relevant entry. This phase, too, might be supported by the exploitation of linguistic data both in the source language and the target language.
The third step of the dictionary building process is the *synthesis stage*, where the final entry will be produced through transforming the translated database records into a series of finished entries for a specific bilingual dictionary.

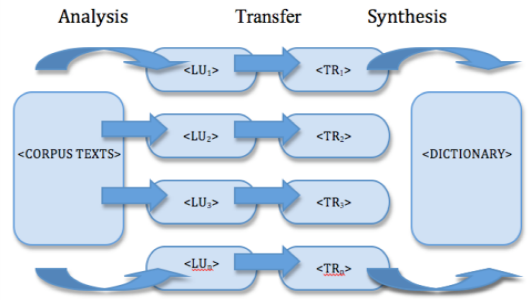

Figure 1: A model of dictionary building

2.3 The Role of Intuition in Dictionary Building

In our days it is widely accepted in the lexicographer community that high-quality dictionaries are based on corpora (e.g. Atkins & Rundell, 2008). The main reason behind this is that linguistic data decreases the role of human intuition during lexicographic process. However, even if lexicographers rely on monolingual data both with respect to the SL and the TL, they inevitably make use of their intuition when deciding which meaningful linguistic units (LUs) have to be included in the dictionary, how to translate them and how to compile the dictionary afterwards.

2.3.1. Intuition in the Analysis Stage

The dictionary building method described above presumes the existence of a monolingual database of senses – LUs – in the SL (a sense inventory), the entries of which will be translated into the TL in the transfer 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. In this subsection two experiments will be shortly presented to emphasize the importance of surface distributional data when creating such monolingual sense inventories.
In the first experiment described in Véronis (2003) 3724 occurrences of 60 words had to be sense-tagged in the context of a one-paragraph text on the basis of the *Petit Larousse* explanatory dictionary by 6 different annotators. The inter-annotator agreement was computed using Cohen's κ. The inter-annotator agreement was relatively low for all the three investigated POS-categories: 0.41 in the case of verbs and adjectives and 0.46 for nouns. Considering the fact that usually 0.8 is accepted as a threshold for reliable agreement (see Artstein & Poesio, 2008), the obtained values imply that Petit Larousse senses are not suitable for the sense-tagging of tokens in their contexts.
A similar experiment was carried out for Hungarian verbs (Héja et al, 2009) based on two different sense inventories, yielding approximately the same conclusion.
One of the sense inventories used was the *Hungarian Explanatory Dictionary*, which is the official reference work as a Hungarian monolingual dictionary.
Another sense inventory tested was the *Hungarian WordNet* (HuWN) (Miháltz et al, 2008). HuWN is a lexical database, modeled partly upon the Princeton WordNet 2.0 (PWN) for English. The basic unit of HuWN, as of all wordnets, is a concept (called synset) and not that of traditional dictionaries, i.e. a lexeme. It is important to note that when deciding on what verb senses should be incorporated into the Hungarian verbal WordNet, automatically extracted information about argument structures was taken into account, as well. Therefore, the sense distinctions in HuWN are partially based on distributional information. Five different annotators sense-tagged the verbs in the context of one sentence in both cases. The inter-annotator agreement was determined using Fleiss's multi π (Artstein & Poesio, 2008). The avarage Fleiss's multi π was 0.3 in the case of the Hungarian Explanatory Dictionary and 0.483 when HuWN was used as sense inventory. Thus, the order of the inter-annotator agreement value was comparable to Véronis' results in the case of both databases. These results clearly show that none of these sense inventories can be exploited to find reliably the relevant meanings of

headwords in 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 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 that exploit linguistic information as much as possible is rather expensive, this approach is typically not affordable in the case of lesser-used languages.

Nevertheless, in the framework of the proposed technique the inter-annotator agreement may be interpreted as measuring the agreement of translators of corpus texts. That is, automatically attained translation-candidates show how frequently a TL expression is assigned to the SL expression, thus capable of indicating the commonly used, recurrent translations.

2.3.2. Intuition in the Transfer Stage

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.

According to Atkins & Rundell (2008: 473) a "*TL corpus has immense potential for dictionary translators"* since *"it offers a way of finding translations, of checking those you are doubtful about, and of correcting those that are simply wrong"*. On the other hand, if representative, a parallel corpus yields a more direct solution to the task of eliminating the role of intuition when hunting for the best translations. Through assigning translation probabilities to translation candidates, the proposed technique is able to estimate which translation is the most frequently used.

Moreover, since all the contexts in which translation candidates occur are directly available, relevant contextual information in the TL side can be extracted to supply information on the proper use of a TL expression. This kind of data is essential in the case of encoding dictionaries.

2.4 Advantages of Parallel Corpora

Beside cost efficiency, one principal advantage of the proposed technique is that it helps to further diminish the role of human intuition. 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 determined 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 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.

The Hungarian-Lithuanian sample entry of *to be born* below illustrates how natural contexts from corpora can help in distinguishing different sub-senses of a word.

HUN LEMMA	LIT LEMMA	TRANSLATIONAL PROBABILITY	FREQUENCY OF HUN LEMMA	FREQUENCY OF LIT LEMMA
Születik	Gimti (-sta,-ė)	0.579005	169	174

HUN	LIT
Ő **1870-ben született**	Jis **gimė 1870 metais**
He **was born in 1870**	
De Fache mintha **erre született** volna	Bet Fasas, regis, tiesiog **tam gimęs**
As if Fache **was born to** do this	
Úgy látszik , **szerencsétlen csillagzat alatt születtél**	Turbūt **gimei po nelaiminga žvaigžde**
It seems that **you were born under an unlucky star**	
…, mert **ikrei születtek.**	…, nes jai **gimė dvynukai.**
…, because **twins were born to her.**	
Maga úriembernek **született.**	Tu **gimei džentlemanu.**
You **was born a gentleman.**	
… hogy Buddha nem **lótuszvirágból született?**	…,kad Buda **gimė ne iš lotoso žiedo?**
…that Buddha **was born from a lotus flower?**	

Figure 2: Sample entry

However, in this example manual lexicographic work is needed to tell the different sub-senses of the Hungarian and Lithuanian counterparts of *to be born* apart. It is important to note that the proposed method allow for the automatic identification of word senses of SL expressions only if different translations are accessible in the TL corpus. Hence, in certain situations the various meanings of words cannot be attained automatically, even in the case of completely unrelated senses (e.g. the German counterpart of the English word *nail* is *der Nagel*, regardless if it denotes the bodypart or the thin pointed peace of metal). Nevertheless, ignoring such cases does not pose a problem for bilingual dictionaries, since several such dictionaries follow the same practice. (e.g. *Collins-Robert French Dictionary* (2006) entry *column*)

2.5 Difficulties

However, besides the essential improvements the proposed method can contribute to traditional or corpus-based lexicography, there are certain difficulties that we have to overcome to be able to create full-fledged core dictionaries of a suitable size.

At this stage of research the proposed method is not capable of handling any kind of multiword expressions i.e. idioms, names, collocations and verbal constructions. Although, based on the provided parallel sentences manual lexicographic work is able to compensate for this shortcoming, the automatic treatment of such expressions is definetely one of our medium-term objectives.

As will be described in 3.1.1 in more detail, the main bottleneck of the method is the scarcity of parallel texts available for medium-density language pairs, due to which the production of an appropriate-size parallel corpus proved to be rather tediuos. Hopefully, with the escalating number of texts accessible in electronic format this task will become increasingly straightforward in the future.

In the next section the construction and evaluation of the Hungarian-Slovenian and Hungarian-Lithuanian core dictionaries will be presented.

3. Workflow

The workflow comprised three main stages. First, resources and language-specific tools had to be collected to create the parallel corpora [3.1]. Secondly, word alignment was carried out to generate the core dictionaries. Based on the preliminary manual evaluation of the Hungarian-Slovenian core 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 [3.2]. Finally, a more precise evaluation of the Hungarian-Lithuanian core dictionary was carried out manually by bilingual speakers, based on criteria that were also defined in this phase [3.3].

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[1] we decided to gather texts translated from a third language. Although national digital archives such as the Digital Academy of Literature[2] and the Hungarian Electronic Library[3] 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 (i.e. tokenization, sentence segmentation and lemmatization – with disambiguation)

[1] For Lithuanian and Hungarian we did not found significant amount of direct translations available in electronic form. In the case of Slovenian and Hungarian, we managed to gather a cc. 750.000-token corpus for each language through contacting several translators, publishers and the Slovenian Television.

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

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

were accomplished by the 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[4] (Erjavec et al., 2005). Hungarian annotation was provided by the pos-tagger of the Research Institute for Linguistics, HAS (Oravecz and Dienes, 2002).

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.

Since our basic goal is to investigate how core dictionaries can facilitate lexicographic process, we sought to minimalize the possible side effects of mismatched sentences. Therefore, corpus texts were manually checked in order to get rid of untranslated sections. Afterwards, a sample of the Hungarian-Slovenian parallel corpus was manually evaluated. Based on the result of the evaluation a threshold had been set and all the aligned sentences with a confidence value below this threshold were discarded in the rest of our analysis. As a result, we have produced two parallel corpora of different sizes. Figure 3. shows the corpus size for each of the language pairs. The 2nd 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

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

Figure 3: Size of the parallel corpora

3.2 Core Dictionaries

This subsection presents how the list of translation candidates was generated [3.2.1], and how the most likely translation candidates were selected to produce the core dictionaries [3.2.2]. In 3.2.3 the evaluation method and the results are described.

3.2.1. Creation of Core Dictionaries

The creation of core dictionaries follows two main steps.

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

The first step is word alignment for which the freely available tool GIZA++ (Och and 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. The retrieved lemma pairs with their translational probabilities served as the starting point for the core 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 evaluation of the Hungarian-Slovenian core 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 core dictionary and a more detailed evaluation was performed on it.

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. This phenomenon is illustrated with two examples in the table below:

HUNGARIAN LEMMA (SL)	HUN FREQ	LITHUANIAN LEMMA (TL)	LIT FREQ	$P(W_t\|W_s)$
arcizom (muscle in the cheeks)	5	jis (he, him, it)	60667	0.852353
ádáz (grim)	23	su (with)	8562	0.797146

Figure 4: Incorrect candidates with high translational probabilities

To filter out such cases an additional constraint was introduced for the Hungarian-Lithuanian language pair: translation candidates where one of the members occurs at least 100 times more than the other were ignored.

3.2.2. Setting the Parameters

The evaluation of a sample Hungarian-Slovenian core 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. As is described above, in the case of Hungarian-Lithuanian a further constraint was added: 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 5 indicates the number of translation candidates that correspond to the parameters determined through the preliminary evaluation. The second column of the table shows the number of expected correct translations, assuming that 65% of the translation candidates with the corresponding parameters are correct.

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

Figure 5: Expected size of the core dictionaries

Considering the fact that we do not intend to create perfect dictionaries, but core dictionaries facilitating lexicographers' work, it seems reasonable to target this value (65%), since it is much easier to throw out incorrect translations than make up new ones. Based on these parameters a detailed manual evaluation of the core Hungarian-Lithuanian dictionary was performed.

Unfortunately, the obtained numbers of expected translation candidates stay far below the targeted size of a medium-sized dictionary (20,000-45,000 entries). Hence, the augmentation of parallel corpora and the refinement of parameters will be definitely part of our future work. The latter is motivated by the fact that many translation candidates with higher frequency proved to be correct translational equivalents, even in the presence of translation probabilities which are at least an order of magnitude lower then the value determined above.

3.3 Detailed Evaluation of the Hungarian-Lithuanian Core Dictionary

The evaluation was performed manually by bilingual (Lithuanian and Hungarian) 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.[5]

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*)

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 core dictionary. All the translation candidates with translational probability 1 (63 pairs) were also included in the evaluation. Figure 6 indicates the result of the evaluation.

[5]Translation candidates are boldfaced in the examples.

	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 %	13 %	49 %	0 %
Sum	Σ 51%		Σ 49%	

Figure 6: Results of the Hungarian-Lithuanian core 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.
Based on this evaluation of the sample, we might expect that 3549 translation candidates out of 4025 should be useful, which yields a better coverage than our original hypothesis (figure 5). Despite the improved results, the coverage of our core dictionary has to be further augmented. One possibility is the refinement of the parameter settings, since the translational probabilities assigned to several correct translation-pairs with higher lemma frequencies are at least an order of magnitude lower than the one determined above.

4 Treatment of Multiple Meanings

As it was pointed out earlier in section 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.
In this section we present two examples to illustrate the above statements.
Taking the claim that *„there is a strong correlation between a word's ferquency and its [semantic] complexity"* (Atkins & Rundell, 2008: 61) as our starting point, we concentrated on cases where Lithuanian lemmas occur at least 100 times in the corpus. In parallel with the augmentation of frequency, we decreased the threshold of translational probability: we set it to 0.02 instead of 0.5. With these parameters we obtained 6500 translation candidates for 1759 Lithuanian lemmas.

4.1 Example 1: *Puikus*

Figure 7 illustrates that the proposed method is able to extract various translations ranked according their likelihood. The translation candidates below support our hypothesis: in the case of more frequent words, translation candidates even with lower probabilities might yield correct results.

LIT	HUN	$P(w_t\|w_s)$	ENG
puikus	jó	0.128	good
puikus	remek	0.071	great, all right
puikus	tökéletes	0.052	perfect
puikus	szép	0.048	nice
puikus	pompás	0.035	splendid
puikus	jól	0.035	well
puikus	nagyszerű	0.035	great
puikus	finom	0.028	fine
puikus	gyönyörű	0.02	marvelous

Figure 7: Hungarian equivalents of *puikus*

The order of the translation candidates might be stunning at first sight for someone who speaks Hungarian, for *remek* which turned out to be the second most probable translation of the Lithuanian *puikus*, is stylistically marked when it modifies a noun. However, the provided examples account for this oddity. In one third of the examples *remek* occurs as a one-word response, which form is quite extensively used in Hungarian. (e.g. -***Puiku***, - atsakė balsas. -**Remek** – válaszolta a hang. (-**All right** – the voice answered)

4.2 Example 2: *Aiškiai*

As it was discussed earlier, the proposed technique seems to be particularly apt to support the creation of encoding dictionaries. If multiple translations are present, it is essential that the choice among them be guided by explicit linguistic criteria. The provided parallel data could be of great help for lexicographers in describing the relevant conditions under which a target language expression could occur. Example 2 illustrates the role of the context in finding the right translational equivalent:

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 terms, 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.

5 Conclusions and Future Work

In this paper a corpus-driven technique was introduced for the purpose of supporting the creation of dictonaries for human use. The proposed automatic method proved useful in supporting such work for several reasons. Most importantly, this approach ensures that the most relevant translations are included in the resulting dictionaries, if representative corpora are available. Moreover, possible translation candidates can be ranked based on their translational probabilities, thus guaranteeing that the most likely translational equivalents go first. Thirdly, all the relevant example sentences are easily accessible, which is of great help in the creation of encoding dictionaries, since these examples could be used as contextual anchors enabling to find the relevant translation in the case of related meanings. Finally, the proposed method renders the generation of the reversed dictionary more straightforward, since solely the word alignment has to be re-applied in the opposite direction.

However, one principal bottleneck of the approach is that the construction of parallel corpora is tedious for medium-density languages. Accordingly, augmentation of the size of our parallel corpora is essential. Further refinement of the parameters also has to be carried out to increase the coverage of the core dictionary.

A further difficulty is that the technique in its present form is unable to handle multiword expressions. One possible solution would be to manually add the missing parts of the expressions based on the provided parallel sentences. Nevertheless, since the automatic treatment of such expressions is highly desirable, this is our future thread of research.

6 Acknowledgements

This pilot project was founded by EFNIL. Our thanks go to Beatrix Tölgyesi and Justina Lukaseviciute for the evaluation of the Hungarian-Lithuanian core dictionary. We are also grateful to Bence Sárossy and Iván Mittelholcz for their contribution to the collection and manual check of the texts.

7 References

Artstein, R.; Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics,* 34(4), pp. 555--596.

Atkins, B. T. S.; Rundell, M. (2008). The Oxford Guide to Practical Lexicography. Oxford: Oxford University Press

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

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. 32--36.

Héja E, Kuti J, Sass B.: Jelentésegyértelműsítés - egyértelmű jelentésítés? In: Tanács A., Szauter D., Vincze V. (eds): *Proceedings of MSZNY2009*, SZTE, Szeged, 2009., p. 348-352.

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, *International Journal of Lexicography*, September 1, 20(3): 221--237.

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

Och, F. J.; Ney, H. (2003). "A Systematic Comparison of Various Statistical Alignment Models". Computational Linguistics, 29 (1). 19--51.

Oravecz, Cs.; 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.; A. Utka. (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. és McEnery, T. (Ed.) Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech. Frankfurt: Peter Lang

Wu, D. (1994), Learning an English-Chinese Lexicon from a Parallel Corpus. In: Proceedings of AMTA'94. 206--213.