Dictionary Building based on Parallel Corpora and Word Alignment

Enikö Héja
Research Institute for Linguistics, HAS, Dept. of Language Technology

The paper describes an approach based on word alignment on parallel corpora, which aims at facilitating the lexicographic work during dictionary building. This corpus-driven technique, in particular the exploitation of parallel corpora, proved to be helpful in the creation of bilingual 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 their translational probabilities it is possible to rank translation candidates, which ensures that the most likely translation candidates are ranked higher. A further advantage is that all the relevant example sentences from the parallel corpora are easily accessible, thus alleviating 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 describes the work accomplished as part of the project EFNILEX[1] founded by the European Federation of National Institutions for Language (EFNIL). The objective of this pilot project was to investigate how language technology methods, in particular parallel corpora can contribute to the 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 work described below aims to produce medium-sized dictionaries for human use covering everyday language vocabulary for Lithuanian and Hungarian. We used Slovenian and Hungarian as a test language pair.

According to the state of the art there are no methods that could enable the fully automatic production of dictionaries. Thus, the creation of a completely clean lexicographical resource with an appropriate coverage requires a manual post-editing phase. Accordingly, our goal is to provide lexicographers with resources diminishing as much as possible the amount of labour required to prepare full-fledged dictionaries for human use. In this paper we will investigate to what extent automatically created translation pairs with their natural contexts are apt for this purpose. Automatically generated resources containing of this information will be referred to as core-dictionaries afterwards.

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 coverage of dictionaries for machine translation purposes, as far as we know, parallel corpora and word alignment has not been exploited in lexicographical projects[2] until now.

The next section shortly presents the advantages and difficulties of relying on sentence-aligned corpora while preparing dictionaries. The 3rd section introduces the workflow itself, the creation of the parallel corpora [3.1] and the core-dictionaries [3.2]. It also describes the

[1] http://www.efnil.org/projects/efnilex

[2] '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' (Atkins and Rundell 2008:447).

evaluation method we used and presents the results of the Hungarian-Lithuanian core-dictionary [3.3]. The 4th section illustrates how the proposed approach copes with multiple meanings. The last section summarizes the conclusions and the remaining tasks [5].

2. Advantages of parallel corpora in dictionary creation

In our days it is widely accepted in the lexicographer community that high-quality dictionaries are based on corpora (e.g. Atkins and 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 linguistic data both on the source language and on the target language side, 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.

Beside cost efficiency, one principle 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 source language and in target corpus and their translational equivalents provided by the parallel sentences. On top of that, 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 possible to determine which sense of a given lemma is the most frequently used. Thus, representative corpora guarantees that not only the most important source lemmata 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 could 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.

At 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 give a detailed description on the use of the target language expression in grammatical terms on the other. Hence, the great amount of easily accessible natural contexts alleviates the creation of encoding dictionaries, when the user needs information on how to translate and use a given expression in a foreign langugae.

Secondly, different subsenses of a headword can be characterized on the basis of the supplied contexts, providing positive evidence that all of these subsenses are translated with the same lemma to the target language.

The Hungarian-Lithuanian sample entry of *to be born* below illustrates how natural contexts from corpora can help in distinguishing different subsenses 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 1. Sample entry from the Hungarian-Lithuanian core-dictionary

However, beside 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 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, thus the production of an appropriate-size parallel corpus proved to be rather tediuos. Hopefully, with the increasing number of texts accessible in electronic format this task becomes 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. At 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 categories that were also defined in this phase [3.3]. Figure 2. gives a more exhaustive description of the workflow. However, only the three main steps mentioned above will be presented comprehensively in the rest of this paper.

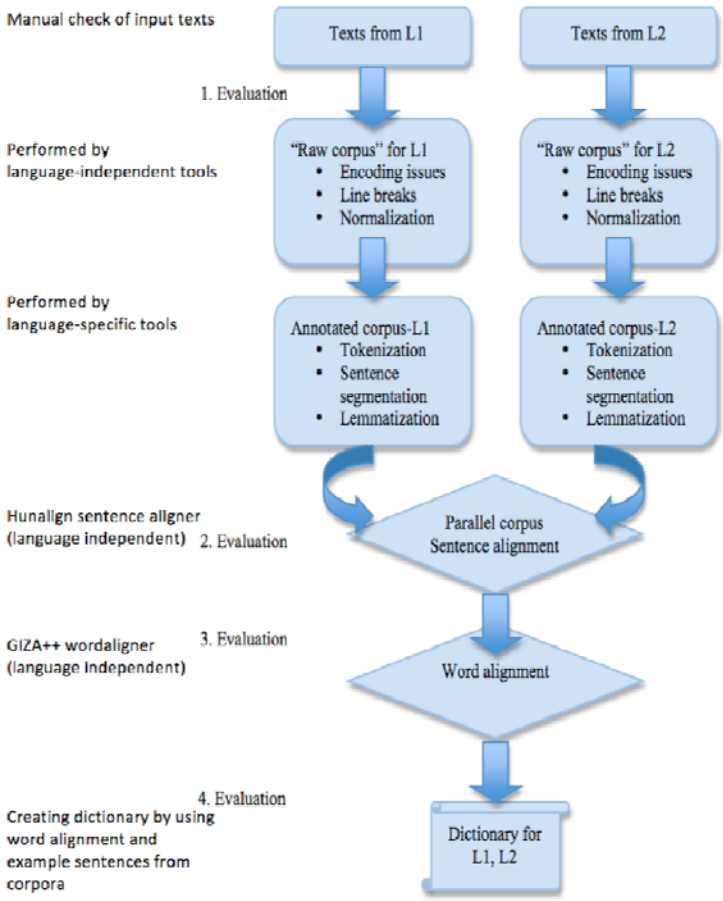

Figure 2. The workflow

3.1. Creation of parallel corpora

3.1.1. Collection of texts and language-specific 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[3] we decided to gather texts translated from a third language. Although national digital archives such as Digital Academy of Literature[4] and Hungarian Electronic Library[5] do exist in Hungary providing us with a wealth of electronically available texts, similar resources has not been found, either for

[3]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.

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

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

Slovenian or 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 the Lithuanian-English parallel corpus (Rimkutė et al. 2008).

As figure 3. shows, basic text-processing tasks (i.e. tokenization, sentence segmentation and lemmatization – with disambiguation) were accomplished by the means of language-specific tools accessible for all these three languages. These tools were available in the form of tool-chains. As for Lithuanian, the analysis was carried out by the Lithuanian Centre of Computational Linguistics (Vytautas Magnus University). Slovanian texts were processed with the tool-chain available at the site of Jožef Stefan Institute[6] (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 as much as possible the problem of data sparseness resulting from rich morphology.

Since our basic objective is to investigate how core-dictionaries can facilitate the lexicographic process, we sought to minimalize the possible side-effects of mismatched sentences. Therefore, corpus texts were at first manually checked 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[7]. 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. Thus, the relevant measure of corpus size is the number of aligned units or translational units, containing at least one sentence in both of the languages.

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

6 http://nl.ijs.si/jos/analyse.

7 Our thanks go to Bence Sárossy and Iván Mittelholcz for their contribution to the collection and manual check of the texts.

3.2. Core-dictionaries

This subsection presents how the list of translation candidates was generated [3.2.1], 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 will be described.

3.2.1. Creation of the core-dictionaries

The creation of core-dictionaries is made up of two main steps. The first step is word alignment for which the freely available word aligner GIZA++ (Och and Ney 2003) was used.

To perform word alignment GIZA++ assigns translational probabilities to source language and target language 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 the 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*. We used Hungarian and Slovenian as a test language-pair, that is, the above parameters were set through the creation of the Hungarian-Slovenian list of translation candidates and used in the evaluation of the Hungarian-Lithuanian core-dictionary. The rationale behind this is quite practical: the availabilty of bilingual speakers for evaluation purposes was rather limited.

The 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 target language lemmata 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 sentence pairs. 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 translation candidates with high translational probablilites

Here SL and TL stand for source language and target language, respectively. 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 discarded in the rest of the analysis.

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 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 the lemma frequency of either the Lithuanian or the Hungarian lemma occurred more than 100 times more 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 as much as possible, 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, these figures 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 parts 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 a 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[8]. 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. too 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

[8] Our thanks go to Beatrix Tölgyesi and Justina Lukaseviciute for the evaluation of the Hungarian-Lithuanian core-dictionary.

expressions or collocations 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 were made up of two subclasses.

In the case of *completely correct translation* pairs no post-editing is needed. In the examples below translation candidates are bolded.

Example 1: HUN: ***gyümölcs*** LIT: ***vaisius*** (*fruit*)

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 matching in the case of a compound, whereas example 3. gives an instance of partial match owing to collocates.

Example 2: (compounds) HUN: ***főfelügyelő*** LIT: *vyriausiasis* ***inspektorius*** (*chief inspector*)

Example 3: (collocations) HUN: ***bíborosi*** testület (*cardinal college*) LIT: Kardinolų ***kolegiją***

Partially correct translations might be the results of slightly loose translations where strong synonymy does not hold between the translation candidates. However, synonymy in the strict sense is quite rare across languages and is mostly confined to quite tight semantic classes e. g. ‘names of concrete objects that the two cultures share’ (Atkins and Rundell 2008:135). Thus, members of this class might yield quite useful clues on how source language and target language lemmata with related meanings can be substituted in certain contexts. Example 4 illustrates the semantic relation of hypernymy.

Example 4: HUN: ***lúdtoll*** (literally: *goose-feather*) LIT: ***plunksna*** (literally: *feather*, *pen*) (intended meaning in both cases: *quill pen*)

Useless translation candidates were made up of *irrelevant translational equivalents*, usually due to recurrent, but irrelevant proper names, exemplified below:

Example 5: HUN: ***Abdul*** LIT: ***Abdulas***

The other subclass of useless translation candidates comprised *completely bad translations*, primarily resulting from too loose translations of texts.

3.3.2. Evaluation methodology and results

Out of the 4025 translation candidates with the parameters determined above[9] 863 pairs were manually evaluated. Throughout the evaluation three intervals were distinguished based on

9 Both lemmata should occur at least five times, and the translational probability equals to or greater than 0.5.

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.

	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 evaluation 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 peliminary estimitation (figure 5). Despite of the improved result, the coverage of our core-dictionary has to be further augmented. One possibility is the refinement of the parameter setting, as several correct translation-pairs with higher lemma frequencies are assigned at least a magnitude lower translational probabilities 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. On top of that, 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 or corpus-based lexicography.

To give a closer look to the above statements, we built also an opposite-direction Lithuanian-Hungarian core-dictionary to be able to compare it with an existing Lithuanian-Hungarian dictionary (Bojtár 2007). Since the comparison has not been carried out completely until now, I will only use this dictionary to illustrate the advantage of this approach over traditional lexicography.

Taking the claim that '*... there is a strong correlation between a word's ferquency and its [semantic] complexity*' (Atkins and Rundell 2008:61) as our starting point, we focused on high-frequency Lithuanian lemmata throughout our analysis, that is, we concentrated on cases where the Lithuanian lemma occurs 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 lemmata.

4.1. Example 1. - *Puikus*

Figure 7. illustrates that the proposed method is apt to extract several diverse translations ranked according their likeliness. The translation candidates listed 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 the Lithuanian word *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. 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.

-**Puiku**, - atsakė balsas. -**Remek** – válaszolta a hang. (-**All right** – the voice answered)

4.2. Example 2. – *Aiškiai*

The proposed technique seems to be particularly apt to support the creation of active or encoding dictionaries (where the native speaker of the source language intends to make utterances in the foreign target language). If multiple translations are present, it is essential that the choice among them be guided by explicit linguistic criteria, be it lexical, grammatical or pragmatical. The provided parallel data could be of great help for lexicographers in

describing the relevant conditions under which a target language expression could be used properly. Figure 8. illustrates the role of contexts in finding the right translational equivalent:

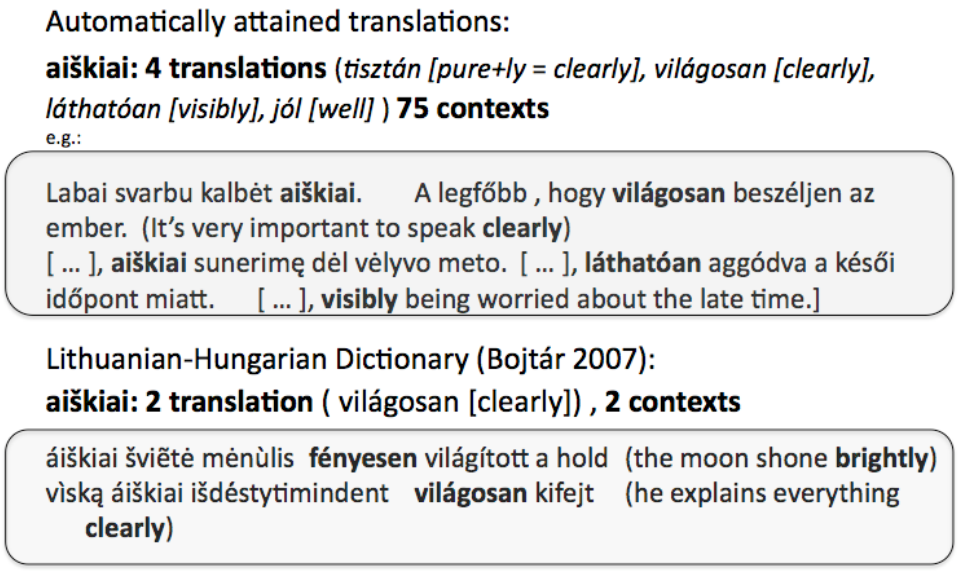

Figure 8. Usage-based, representative translations

Whereas in the traditional dictionary 2 translations are included with 2 different contexts, our core-dictionary lists 4[10] translations along with 75 contexts.

5. Conclusion 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 to be capable to aid such works for several reasons. Most importantly, this approach ensures that – if reperesentative parallel corpora are available – most relevant translations are included in the resulting dictionaries. On top of that, possible translation candidates can be ranked based on their translational probabilities, thus guaranteeing that 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 when picking out the relevant translation in the case of related meanings. Finally, the proposed method renders the generation of the reversed direction dictionary more simple, since solely the word alignment has to be re-applied in the opposite direction.

However, one principle bottleneck of this approach is that the construction of parallel corpora is a quite time-consuming task for medium-density languages. Accordingly, one of our main tasks is the augmentation of the size of our parallel corpora. Further refinement of the parameters also has to be carried out to augment the coverage of the core-dictionary.

An other difficulty resulting from the word alignment algorithm 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.

[10] Our core-dictionary comprised 6 translation candidates corresponding to the parameters, out of which 4 were correct translations.

References

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

Digitális Irodalmi Akadémia [Digital Academy of Literature]. (2007) *Litván-magyar nagyszótár* [Lithuanian-Hungarian Dictionary]. Budapest: Akadémiai kiadó. 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.

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

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.

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