Utilizing Citations of Foreign Words in Corpus-Based Dictionary Generation

Reinhard Rapp
University of Tarragona
GRLMC
reinhardrapp@gmx.de

Michael Zock
Laboratoire d'Informatique Fondamentale
CNRS Marseille
Michael.Zock@lif.univ-mrs.fr

Abstract

Previous work concerned with the identification of word translations from text collections has been either based on parallel or on comparable corpora of the respective languages. In the case of comparable corpora basic dictionaries have been necessary to form a bridge between the languages under consideration. We present here a novel approach to identify word translations from a single monolingual corpus without necessarily requiring dictionaries, although, as will be shown, a dictionary can still be useful for improving the results. Our approach is based on the observation that for various reasons monolingual corpora typically contain many foreign words (for example citations). Relying on standard newsticker texts, we will show that their co-occurrence-based associations can be successfully used to identify word translations.

1 Introduction

The web has popularized information access. As a consequence, the information put on the web evolved, expanding from mainly technical documents in one language (English) to topics concerning nearly any aspect of life in many languages. For this reason it cannot be expected anymore that all web users speak English. Yet users speaking only one of the minority languages will be penalized, finding only a small fraction of web content accessible. Hence they can make only very limited use of what is available. In order to increase information access independently of the users' mother tongue, automatic translation is desirable.

Recognizing this need, Google, among others, is providing free machine translation services for any pair of currently 50 languages.[1] However, with 6800 living languages, of which 600 also use a written form, offering comprehensive translation services remains a challenge.

The statistical approach to machine translation (SMT), as adopted by Google, relies on parallel corpora, i.e. large collections of existing translations. But it is a daunting task trying to acquire parallel corpora for all possible language pairs. Therefore, it appears that for some languages Google has combined SMT with an interlingua approach. This allows optimal exploitation of languages for which parallel corpora are easily obtained. These languages are then used as pivots. Note that in phrase-based SMT an interlingua approach may operate at the level of the phrase table, which facilitates matters while speeding up the process. At the downside it must be noted that a phrase table derived via a pivot language is generally of lower quality than a phrase table directly compiled from parallel texts (provided the corpus size is similar). Hence, just as for other interlingua approaches, translation quality is severely compromised.

An alternative approach that has been suggested is to try to generate the required dictionaries from other sources than parallel corpora. Bear in mind that statistical machine translation requires a *language model* and a *translation model*. To generate the language model only monolingual corpora of the target language are required which, for example, can be acquired from the web. If only few such documents exist, one may well conclude that there is probably no real need

[1] http://www.google.de/language_tools?hl=de as of April 22, 2010.

Proceedings of the Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010), pages 50–59,
Beijing, August 2010

for translation involving this particular language. So the main bottleneck are the parallel corpora required to generate a translation model. But the purpose of the translation model is in essence the creation of a bilingual dictionary, be it a dictionary of individual words or a dictionary of phrases. For this reason, if we can find other ways to generate dictionaries for lesser used languages, this will be beneficial not only for the users of these languages but also for the solution of the overall problem of machine translation.

In other words, an important challenge is the generation of dictionaries. Since comparable corpora are a far more common resource than parallel corpora, attempts to exploit them for dictionary construction have received considerable attention recently.[2]

One approach is to mine parallel sentences from comparable corpora. Roughly speaking, this can be done by automatically translating a corpus from one language (source language) to another (target language), and then searching in a large corpus of the target language for sentences similar to the translations. The advantage of this procedure is that the sentences retrieved this way are correct sentences as they were produced by humans, whereas the sentences translated by a machine tend to be garbled and of lower quality. However, the big problem with this approach is to ensure that the retrieved sentence pairs are indeed translations of each other. While there is no perfect solution to this problem, several studies have shown that such data can be useful for building or supplementing translation models in SMT (see e. g. Munteanu & Marcu, 2005; Wu & Fung, 2005).

Another approach for exploiting comparable corpora in dictionary generation is based on the observation that word co-occurrence patterns between languages tend to be similar (Fung & McKeown, 1997; Rapp, 1995; Chiao et al., 2004). If, for example, two words X and Y co-occur more often than expected by chance in a corpus of language A, then their translated equivalents should also co-occur more frequently than expected in a corpus of language B. A great number of variants of this approach has been proposed, e.g. emphasizing aspects of corpus selection or expanding it to collocations or short phrases (Babych et al., 2007).

What is common to these studies is that they consider the source and the target language as two distinct semantic spaces, without any links at the beginning. Therefore, in order to connect the two, a base dictionary is required, and the purpose of the system is to expand this base dictionary. Building a dictionary from scratch is not possible this way or at least computationally unfeasible (see Rapp, 1995).

Whether the assumption of two completely distinct semantic spaces is realistic remains an open issue. Are separate lexical networks really a reasonable model for the processing of different languages by people?

One could say this is a plausible model, assuming a person lived for some years in one country, and then for some more years in another country, assuming further that this person never looked at a dictionary or another multilingual document and never communicated with a person mixing both languages.

It is known that this can work. The reason is probably the following: Many words of the basic dictionary assumed above correspond to items of the physical world. These items generally have names in natural languages which can serve as mediators. That the extrapolation to more abstract notions is possible has been claimed by Rapp (1999).

Still, although persons proceeding this way can easily understand and, after some years, even think in each of the two languages, experience shows that they tend to have some difficulties when making translations, especially literal translations.

So, although the above scenario is possible, we do not think that it is a typical one for our modern times. There are certainly good reasons why there are so many language courses, and why there is such an abundance of dictionaries. It is a matter of commonsense that the person trying to acquire a new language will look at a multilingual dictionary. He or she will also communicate with other persons who mix languages, for example, relatives, other people from the com-

[2] There is also the approach of identifying orthographically similar words (Koehn & Knight, 2002) which does not even require a corpus as simple word lists will suffice. However, this approach is promising only for closely related languages but appears to have limited scope otherwise. For this reason we will not further discuss it here.

munity of foreigners coming from the same country, teachers in language classes, etc. In many cases there will also be multilingual documents around: leaflets, explanations in a museum, or signs in a public area (e.g. airport).

Hence the spoken and written "corpus" (input) on which such a person's language acquisition process is based is not solely monolingual. While the corpus may be mainly monolingual, it surely will contain some multilingual elements.

If we agree on this, our next step could be to acquire transcripts of language teaching classes with bilingual teachers and try to exploit these for dictionary generation. Since obtaining such transcripts in large enough quantities should be much more difficult than obtaining parallel corpora, this approach will probably not solve the data acquisition bottleneck which is the practical problem we were about to solve in the first place.

The current study is therefore based on newsticker texts which is a text type very similar to standard newspaper texts. At least for some languages it is available in large quantities. However, this type of text is probably not ideally suited for our purpose. Surprisingly, the reason is that newsticker and newspaper texts tend to be very well edited. This means that the author will typically avoid foreign words, and if ever some remain the respective passages are likely be rephrased in order to make sure that the text uses familiar vocabulary, easily understandable by the readers. However, this is problematic for our approach which is based on the occurrences of foreign words in a monolingual text. So this is one of the rare cases where noisy corpora should yield better results than perfectly clean data.

On the other hand, as this study suggests a (to our knowledge) novel approach, we consider it important to use a corpus that is generally known and available, and which has not been compiled with this particular purpose in mind. Only this way our results can convincingly give an idea concerning the baseline performance of the suggested algorithm. At this stage we consider this more important than optimizing results by compiling corpora specifically suited for the purpose, even though this will be a logical next step.

2 Approach and Language Resources

Starting from the observation that monolingual dictionaries typically include a large number of foreign words, we consider the most significant co-occurrences of them as potential translation candidates. This implies that the underlying corpus corresponds to the target language, and that it can be utilized for any source language for which it contains a sufficient number of word citations. As this paper is written in English, we chose an English corpus as this should make judging our results convenient for most readers. However, being the world's most widely spoken language, English tends to be rather self-contained in comparison to other languages, which may use foreign words more frequently. In particular, as a side effect of globalization, the use of English terminology is popular in many other languages. Therefore, in order to identify, for example, German–English word translations, it is better to look at occurrences of English words in a German corpus rather than at occurrences of German words in an English corpus.[3]

Nevertheless, the corpus we use here is the latest release of the English Gigaword Corpus (Fourth Edition) provided by the Linguistic Data Consortium (Parker et al., 2009). It consists of newswire texts of the time between 1995 and 2008 from the following news agencies:

- Agence France-Presse, English Service
- Associated Press Worldstream, English Service
- Central News Agency of Taiwan, English Service
- Los Angeles Times/Washington Post Newswire Service
- New York Times Newswire Service
- Xinhua News Agency, English Service

Altogether, the corpus comprises about 3 billion words. Since we are not interested in the translation of function words, and in order to reduce the computational load, we removed all function words that were included in a stop word list for English comprising about 200 items. The stop words had been manually selected from a corpus-derived list of high frequency words.

In the resulting corpus associations between words need to be identified, something that is usually done on the basis of co-occurrences. In

[3] Note that the results of both directions may be combined. This is something we leave for future work.

order to count the co-occurrences between pairs of words, a text window comprising the ten words preceding and following a given foreign word is considered. On the resulting co-occurrence counts a standard association metric like the log-likelihood ratio (Dunning, 1993) is applied.

Note that the above mentioned window size of ±10 words from the given word relates to the preprocessed corpus from which function words have already been removed. Since in English roughly every second word tends to be a function word, the effective window size is about ±20 words. This window size is somewhat larger than what we typically find in other studies. However, the reason for this is quite obvious: As citations of foreign words are rare, we have a severe problem of data sparseness, and by looking at a relatively large window we try to somewhat compensate for this.[4]

Despite its simplicity, this procedure of computing associations to foreign words already works well for identifying word translations. We simply assume that the strongest association is the best translation. We used this approach for words from three languages: French, German, and Spanish. The results are presented in the next section. In order to measure the quality of our results, for all source words of a language we counted the number of times where the expected English target word obtained the highest association score.

As our gold standard for evaluation we used an existing list of translations as described in Rapp & Zock (2010), i.e. a resource that had not been compiled with the current application in mind. The data consists of 1079 word equations in three languages: English, French, and German. It has been extracted from the respective editions of the Collins GEM dictionaries, whereby when looking up a word only the first entry in the list of possible translations was taken into account. As in the current study we are also interested in Spanish, we manually looked up the main translations at the leo.dict.org website[5] and added another column to this resource. Table 1 shows a few sample entries of the resulting list of *word equations* which were used for evaluating our approach.

We should mention that the term *word equation* is a bit problematic, as most words tend to be ambiguous, and ambiguities tend to vary with language. For this reason, we should, at least in principle, disambiguate all words in our corpus and map them to unambiguous concepts. Next we should use a gold standard using such concepts rather than words. Unfortunately, the current state of the art does not allow doing this with sufficient accuracy. Anyhow, addressing this problem is well beyond the scope of this paper.

SOURCE LANGUAGES			TARGET LANG.
FRENCH	GERMAN	SPANISH	ENGLISH
britannique	britisch	británico	British
Pâques	Ostern	Pascua	Easter
capable	fähig	capaz	able
accent	Akzent	acento	accent
accident	Unfall	accidente	accident
accordéon	Akkordeon	acordeón	accordion
acide	Säure	ácido	acid
gland	Eichel	bellota	acorn
action	Handlung	acción	action
avantage	Vorteil	ventaja	advantage

Table 1. Some sample entries from the gold standard of word equations.

So far, for identifying the translations of the 1079 French words, we assumed the following approach: We first computed their associations and then conducted an evaluation by checking for how many words the top association was identical to the English translation found in the gold standard. The same approach was also used for the other languages, namely German and French. Hence, the three source languages were treated completely independently of each other.

[4] In preliminary experiments we also experimented with other window sizes. However, as we noticed that changes within a reasonable range of e.g. 5 to 20 words have only little effect, we do not consider them here.

[5] This is a manually edited high quality online dictionary. Although it can be used for free, in our view for many purposes is as good as or even better than conventional printed dictionaries.

However, there are several problems with this approach, in particular:

a) Several correct translations
b) Data sparseness
c) Homograph trap

Let us discuss these issues point by point.

a) Several correct translations

Suppose we tried to identify the translation of the German word *Straße* and our gold standard listed *street* as the correct translation. If, however, our system produced *road* this would be considered just as much of an error as if it had produced a very remote word such as *volcano*. Hence, considering only a single word as being correct, which is the consequence of using as gold standard the resource exemplified in Table 1, implies that performance figures are artificially low, giving us only the lower bound of the true performance.

Despite this shortcoming, we will nevertheless do so for the following reasons: 1) This is a pilot study presenting a new approach. For this reason, clarity has priority over performance. 2) The number of translations listed in a dictionary typically depends on the size of the resource. Hence, there is no absolute difference between correct and incorrect translations. Rather, we need to set a threshold somewhere, and truncating after the first word listed is arguably the clearest and simplest way of doing so. 3) This is the main reason. We want to extend our approach to the multilingual case by (simultaneously) looking at several source languages. Given the fact that each language tends to have its own (i.e. idiosyncratic) ambiguities, we are already satisfied if words from the various source languages have the same main translation. That all possible translations are identical is very unlikely.

b) Data sparseness

What will happen if a source word does not occur at all in the corpus, or only once or twice? We mentioned already that an appropriate choice of text genre, corpus size, and window size can somewhat reduce the problem of data sparseness. We also mentioned that by reversing source and target languages we can look at the problem from two perspectives, which may yield further improvement. Nevertheless, these suggestions are limited in scope. Hence, given the nature of our approach, data sparseness will remain the core problem.

Fortunately, there is another possibility which is more promising than the ones mentioned above, provided that we manage to solve the ambiguity problem. The solution consists in considering several source languages concurrently. Suppose that rather than starting from scratch we use existing dictionaries for various languages.[6] In this case we can easily generate word equations such as the ones shown in Table 1. We do this by considering as a single item all words appearing in a given row (excluding the target language word), and by computing the associations to this aggregated artificial unit. (This is a simplified proposal. We shall see later how to improve it.) If, for example, we have 10 source languages, then it does not matter that 8 source words do not occur in the corpus, as long as the other two are well represented.

c) The homograph trap

By this we mean that a word form from the source language also exists in the target language, but with a different meaning. For example, let us assume that we wanted to translate the word *can* (house) from Catalan to English. Suppose further that we are lucky and have ten Catalan citations with this word in our English corpus. But this will not help us because the word *can* happens to also belong to English, meaning something completely different. Moreover, *can* is a high frequency word, occurring millions of times in a large corpus. Of course, if we had a perfect word sense disambiguator, we could separate the Catalan and the English occurrences of *can*, thereby solving the problem.[7] Unfortunately, existing tools are not powerful enough to do the job. What is worse, such collisions are not

[6] Which, for example, by using open source tools such as Moses and Giza++ (see www.statmt.org) can be easily generated from parallel corpora, e.g. from the Europarl corpus (Koehn, 2005) or the JRC Acquis corpus (Steinberger et al., 2006).

[7] If we assume that foreign words typically occur in clusters, we could also use language identification software.

uncommon between languages using the same script. So what can we do? Our suggestion is exactly the same as above for the problem of data sparseness, i.e. to look at several source languages in parallel.

But it is clear that collapsing all source words into a single item does not work. If only one of them happens to be also a common word in the target language, it is very likely that its co-occurrences will override the co-occurrences of the foreign words we are interested in. So there is little chance to come up with a correct result.

We propose a relatively simple solution to this problem, which possibly may well be novel in this context. Let us develop the idea.

In preliminary experiments we have tried several possibilities. Collapsing the source words would be equivalent to adding the respective co-occurrence vectors. This is apparently not adequate because, as mentioned above, the vector of a very frequent word would dominate all others. An alternative would be to sum up the association vectors. By the term association vector we mean the co-occurrence vectors after application of an association measure (in our case the log-likelihood ratio). It turns out that this somewhat reduces the problem without solving it entirely. Another possibility would be vector multiplication. Multiplication is considerably better than addition as a property of multiplication is that moderate but coinciding support for a particular target word from several source words leads to a higher product than strong support by only a few. This is a highly desirable property as it helps us avoiding the homograph trap, and because all values are subject to considerable sampling errors.

Unfortunately, there is yet another problem. Our association measure of choice, namely the log-likelihood ratio, as typical for ratios, has a skewed value characteristic. Since otherwise our previous experiences with the log-likelihood ratio are very good,[8] and since it seems reasonably well suited for sparse data (Dunning, 1993), we suggest to multiply log-likelihood ranks rather than log-likelihood scores. This proposal is based on the observation (Dunning, 1993) that rankings of association strengths as produced by the log-likelihood ratio tend to be highly accurate even at higher ranks. Let us call this procedure the *product-of-rank* algorithm

This algorithm works as follows: Starting from a vocabulary of target language words (which are the translation candidates), for each of these words an association vector is computed. Next, for each association vector the ranks of all words in the source language word tuple under consideration are determined. Hence, if we have three languages (e.g. English, French and German) we would get three values. These values are multiplied with each other, and finally all target language words are sorted according to the resulting products. As small ranks stand for strong associations, the word obtaining the smallest value is considered to be the translation of the source language tuple. This algorithm turned out to lead to highly plausible rankings and to be robust with regard to sampling errors.[9] It is also quite effective in eliminating the homograph problem.

3 Experimental Results and Evaluation

Let us first try to see whether the basic assumption underlying our approach is sound, namely that we will find a sufficient number of foreign words in our corpus. To check this claim, we have listed in Table 2 for each of the four languages the number of words from the gold standard falling into particular frequency categories. For example, the value of 70 in the field belonging to the row *6-10* and the column *Spanish* means that out of the 1079 Spanish words in our gold standard 70 have a corpus frequency between 6 and 10 in the 4th edition of the English Gigaword Corpus. Apparently, words with zero occurrences or with a very low corpus frequency are problematic because of data sparseness. Yet words with very high frequencies are not less problematic, as they may turn out to have homographs in the target language. As there is no generally accepted definition of what the vocabulary of a given language is, we cannot give precise figures concerning the number of homographs in our gold standard for each language pair. Never-

[8] To the best of our knowledge no other measure could consistently beat it over a wide range of NLP applications.

[9] A further improvement is possible by giving words with identical association strengths not arbitrary ranking positions within this group, but an average rank which is to be assigned to all of them.

theless, we believe that Table 2 gives a fair impression. By taking a look at the high frequency source language words one can see that the pair French–English has the greatest number of homographs, followed by German–English, and finally Spanish–English.

Corpus frequency	Source languages			Targ. lang.
	German	French	Spanish	English
0	449	329	317	0
1	64	85	43	0
2	26	52	25	0
3	24	39	23	0
4	17	34	27	0
5	7	26	15	0
6-10	32	71	70	0
11-20	50	59	86	0
21-50	63	52	129	0
51-100	50	37	95	1
101-200	52	10	75	3
201-500	50	25	74	6
501-1000	43	18	31	19
1001-10000	100	71	37	245
above 10000	52	171	32	805

Table 2: Corpus frequencies of the words occurring in the gold standard.

As to be expected, the corpus frequencies of the language of the corpus, namely English, are orders of magnitude higher than those of the other languages. But the table also gives a good idea concerning the presence of French, German, and Spanish word citations in written English. However, we should not be misled by the overwhelming presence of French words in the high frequency ranges, as this mainly reflects the amount of homography. Although pronunciation rules are very different between English and French, spelling tends to be similar, which is why there are lots of homographs. In contrast, Spanish and German usually use different spelling even for words having the same historical roots, which is why homography is far less common.[10]

[10] As an example for such spelling conversions, let's mention that the grapheme *c* in English is almost consistently replaced by *k* in German, e.g. *class* → *Klasse* and *clear* → *klar*.

From the figures of Table 2 one may conclude that identifying word translations from a monolingual corpus is not easy because of data sparseness. Nevertheless it seems possible, at least to some extent. Let us therefore take a look at some results.

In our experimental work we first identified word translations for stimulus words from a single source language, then for stimulus words from two source languages, and finally for stimulus words from three source languages.

a) One source language

We started by conducting separate runs for each of the three source languages (French, German, Spanish) and determined the number of times the algorithm was able to come up with the expected English translation as the top ranked association for the 3 * 1079 source words. Note, however, that hereby we did not consider the full range of possible target words present in the English Gigaword corpus as this would include many foreign words. Instead, we restricted the number of target words to the 1079 English words present in the gold standard.

The respective figures are 163 (15.1%) for French, 85 (7.9%) for German, and 97 (9.0%) for Spanish. As can be seen, French clearly performed best, which confirms previous studies that the lexical agreement between French and English is surprisingly high. Nevertheless, on average, only 10.7% of the translations were identified correctly, which does not look very good. However, remember that these figures can be considered as a lower bound as we do not take alternative translations into account and as the underlying corpus has not been prepared specifically for this purpose. Note also that the *product-of-ranks* algorithm has no effect in the case when only a single source language is considered. (If there is only one value, no multiplication takes place.)

b) Two source languages

Our next step was to combine pairs of source languages. There are three possible pairs, namely French–German, French–Spanish, and German–Spanish. Their respective performance figures are as follows: 217 (21.0%), 225 (20.9%), and 145 (13.4%). Computing the mean of these re-

sults yields an average of 18.4%, which is a nice improvement over the initial 10.7% which we had for single source languages. This lends support to our hypothesis that the product-of-ranks algorithm works effectively in this context.

c) Three source languages

Finally, all three source languages were combined, resulting in the correct translation of 248 of the altogether 1079 test items, which corresponds to a performance of 23.0%. This further improvement is consistent with our hypothesis that performance should increase when more source languages are considered.

Let us take a closer look at these performance gains. At the beginning we increased the number of source languages by 100% (from 1 to 2), yielding a relative performance increase of 72% (the absolute performance improved from 10.7% to 18.4%). Next we increased the number of source languages by 50% (from 2 to 3) which yielded a relative performance increase of 25% (absolute performance had improved from 18.4% to 23%). This means that the behavior is worse than linear, as in the linear case we should have obtained a further improvement of 72%/2 = 36%. But of course when combining statistics in NLP, hardly ever a linear behavior can be observed, and the above findings seem satisfactory. Nevertheless they should be supported by looking at further languages, see Section 4.[11]

For the case of looking at three source languages in parallel, let us provide data concerning the rank distribution of the expected translations (see the middle column of Table 3). Overall, in 357 of the 1079 cases (33.9%) the expected translation ranks among the top five, and in 392 cases (36.3%) it is among the top ten associations. These results are based on a window size of ±10 words when counting the co-occurrence frequencies. To give an idea that the procedure is robust in this respect, we provide analogous values for a window size of ±20 words in the third column of Table 3. As can be seen, apart from the usual statistical fluctuations the difference is hardly noticeable.

Rank	Number of items with the respective rank	
	window size ±10	window size ±20
rank could not be computed (all source words unknown)	11	10
1	248	247
2	55	51
3	32	36
4	15	19
5	7	8
6	16	8
7	7	6
8	3	5
9	3	5
10	6	4
above 10	676	680

Table 3: Ranks of the expected translations when all three source languages are combined.

EXAMPLE 1

Given word French: tablier [7]
Given word German: Schürze [0]
Given word Spanish: delantal [4]

Expected translation into English according to the gold standard: apron [3059]

Top 5 translations as computed:

1	apron	[3059]
2	sausage	[9954]
3	sauce	[49139]
4	appetite	[24682]
5	mustard	[13477]

Table 4: Sample results.

[11] Another important question, which we have not dealt with yet, is to what extend the observed gain in performance when increasing the number of source languages is a side effect of a higher likelihood that at least one of the source words happens to be identical to the target word (with the same or a similar meaning). In such cases (which might be common when considering related languages), predicting the correct translation is rather easy.

EXAMPLE 2

Given word French: carton [2671]
Given word German: Karton [22]
Given word Spanish: cartón [0]

Expected translation into English
according to gold standard: cardboard [13714]

Top 5 translations as computed:

1	cardboard	[13714]
2	cigarette	[54583]
3	fold	[43682]
4	milk	[85426]
5	egg	[42948]

Table 5: Sample results.

Having looked at the quantitative results, some sample output may also be of interest. For this purpose, Tables 4 and 5 show sample results for triplets of source language words. Hereby, the numbers in square brackets refer to the corpus frequencies of the respective words in the English Gigaword Corpus.

4 Summary and Future Work

In this paper we made an attempt to solve the problem of identifying word translations on the basis of a single monolingual corpus where the same corpus is supposed to be used for several language pairs. The basic idea underlying our work is to look at citations of foreign words, to compute their co-occurrence-based associations, and to consider these as translations of the respective words.

We pointed out some difficulties with this approach, namely the problem of data sparseness and the homograph trap, but were able to suggest and implement at least partial solutions. Using the product-of-ranks algorithm, our main suggestion was to look at several source languages in parallel, which at least in theory has the potential to solve the experienced problems.

We did not have very high expectations when starting this work and were positively surprised by the resulting performance of up to 25% correctly predicted test items. As pointed out, in order to avoid raising unjustified expectations, we presented somewhat conservative figures which should leave room for improvements.

Obvious extensions of the current work are to increase the number of considered languages and to also use other large monolingual corpora. For example, we could use the web corpora provided by the web-as-a-corpus (WaCky) initiative (Baroni et al., 2009). A few such corpora have already been made available recently, and as they are based on a largely automatic acquisition procedure there are probably more to come. This reflects a tendency towards extremely large corpora. Processing in the current framework turns out to be unproblematic if sparse matrices are used, as foreign word occurrences are implicitly of low frequency.

Although web corpora should be very noisy in comparison to the carefully edited newsticker texts used here, the interesting thing is that according to the hypothesis formulated in the introduction the current approach seems to provide one of the rare occasions where noisy data is better than perfectly clean data, and we hope that future work will prove this prediction correct.

Another possibility for future work is to look at second rather than first order associations, i.e. to consider those words as potential translations of a given foreign word which show similar context words. This might be promising in so far as the sparse data problem is less salient in this case.

Finally, let us come back to our speculative question from the introduction whether or not people speaking different languages have separate lexico-semantic networks in their mind. Aparently our experiments did not provide evidence for either assumption. But the most straightforward assumption would probably be that our mind does not attach language labels to the words we perceive, and simply treats them all equally. At the lexical level, our mind's unknown inner workings may be in effect analogous to clustering words according to their observed co-occurrence patterns. The likely result is that in some cases there will be many interconnections between clusters, and in other cases few. Depending on the language environment experienced by a person, we cannot rule out that some of the larger clusters might exactly correspond to languages. But what the current research does

tell us is that there can be a multitude of statistically significant co-occurrences even at non-obvious places. So what we possibly should rule out is that, even across languages, there are separate clusters without any interconnections.

Acknowledgments

Part of this research was supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme. We thank the Linguistic Data Consortium for making available the English Gigaword Corpus, and Lourdes Callau, Maria Dolores Jimenez Lopez, and Lilica Voicu for their support in acquiring it.

References

Babych, Bogdan; Sharoff, Serge; Hartley, Anthony; Mudraya, Olga (2007). Assisting Translators in Indirect Lexical Transfer. *Proceedings of the 45th International Conference of the Association for Computational Linguistics ACL 2007, Prague*, 136–143.

Baroni, Marco; Bernardini, Silvia; Ferraresi, Adriano, Zanchetta, Eros (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation* 43 (3): 209–226.

Chiao, Yun-Chuang; Sta, Jean-David; Zweigenbaum, Pierre (2004). A novel approach to improve word translations extraction from non-parallel, comparable corpora. In: *Proceedings of the International Joint Conference on Natural Language Processing*, Hainan, China. AFNLP.

Dunning, Ted (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Fung, P.; McKeown, K. (1997). Finding terminology translations from non-parallel corpora. *Proceedings of the 5th Annual Workshop on Very Large Corpora,* Hong Kong, 192–202.

Fung, P.; Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In: *Proceedings of COLING-ACL 1998,* Montreal, Vol. 1, 414–420.

Koehn, Philipp; Knight, Kevin (2002). Learning a translation lexicon from monolingual corpora. In: Unsupervised Lexical Acquisition. Proceeding of the ACL SIGLEX Workshop, 9–16.

Koehn, Philipp (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of MT Summit*, Phuket, Thailand, 79–86.

Munteanu, Dragos Stefan; Marcu, Daniel (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics* 31(4), 477–504.

Rapp, Reinhard (1995). Identifying word translations in non-parallel texts. In: *Proceedings of the 33rd Meeting of the Association for Computational Linguistics.* Cambridge, Massachusetts, 320–322.

Rapp, Reinhard. (1999). Automatic identification of word translations from unrelated English and German corpora. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics 1999,* College Park, Maryland. 519–526.

Rapp, Reinhard; Zock, Michael (2010). Automatic dictionary expansion using non-parallel corpora. In: Andreas Fink, Berthold Lausen, Wilfried Seidel Alfred Ultsch (Eds.) *Advances in Data Analysis, Data Handling and Business Intelligence. Proceedings of the 32nd Annual Meeting of the GfKl, 2008*. Heidelberg: Springer.

Steinberger, Ralf; Pouliquen, Bruno; Widiger, Anna; Ignat, Camelia; Erjavec, Tomaž; Tufiş, Dan; VARGA, Dániel (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5thLREC***,** Genoa, Italy.

Wu, Dekai; Fung, Pascale (2005). Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. Proceedings of the *Second International Joint Conference on Natural Language Processing (IJCNLP-2005)*. Jeju, Korea.

Parker, Robert, Graff, David; Kong, Junbo; Chen, Ke; Maeda, Kazuaki (2009). *English Gigaword. Fourth Edition.* Linguistic Data Consortium, Philadelphia.