The Noisier the Better: Identifying Multilingual Word Translations Using a Single Monolingual Corpus

Reinhard Rapp
University of Tarragona
GRLMC
reinhardrapp@gmx.de

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

Abstract

The automatic generation of dictionaries from raw text has previously been based on parallel or comparable corpora. Here we describe an approach requiring only a single monolingual corpus to generate bilingual dictionaries for several language pairs. A constraint is that all language pairs have their target language in common, which needs to be the language of the underlying corpus. Our approach is based on the observation that monolingual corpora usually contain a considerable number of foreign words. As these are often explained via translations typically occurring close by, we can identify these translations by looking at the contexts of a foreign word and by computing its strongest associations from these. In this work we focus on the question what results can be expected for 20 language pairs involving five major European languages. We also compare the results for two different types of corpora, namely *newsticker texts* and *web corpora*. Our findings show that results are best if English is the source language, and that noisy web corpora are better suited for this task than well edited newsticker texts.

1 Introduction

Established methods for the identification of word translations are based on *parallel* (Brown et al., 1990) or *comparable corpora* (Fung & McKeown, 1997; Fung & Yee, 1998; Rapp, 1995; Rapp 1999; Chiao et al., 2004). The work using parallel corpora such as Europarl (Koehn, 2005; Armstrong et al., 1998) or JRC Acquis (Steinberger et al., 2006) typically performs a length-based sentence alignment of the translated texts, and then tries to conduct a word alignment within sentence pairs by determining word correspondences that get support from as many sentence pairs as possible. This approach works very well and can easily be put into practice using a number of freely available open source tools such as Moses (Koehn et al., 2007) and Giza++ (Och & Ney, 2003).

However, parallel texts are a scarce resource for many language pairs (Rapp & Martín Vide, 2007), which is why methods based on comparable corpora have come into focus. One approach is to extract parallel sentences from comparable corpora (Munteanu & Marcu, 2005; Wu & Fung, 2005). Another approach relates co-occurrence patterns between languages. Hereby the underlying assumption is that across languages there is a correlation between the co-occurrences of words which are translations of each other. If, for example, in a text of one language two words *A* and *B* co-occur more often than expected by chance, then in a text of another language those words which are the translations of *A* and *B* should also co-occur more frequently than expected.

However, to exploit this observation some bridge needs to be built between the two languages. This can be done via a basic dictionary comprising some essential vocabulary. To put it simply, this kind of dictionary allows a (partial) word-by-word translation from the source to the target language,[1] so that the result can be considered as a pair of monolingual corpora. Deal-

[1] Note that this translation can also be conducted at the level of co-occurrence vectors rather than at the text level.

Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING 2010, pages 16–25,
Beijing, August 2010

ing only with monolingual corpora means that the established methodology for computing similar words (see e.g. Pantel & Lin, 2002), which is based on Harris' (1954) *distributional hypothesis*, can be applied. It turns out that the most similar words between the two corpora effectively identify the translations of words.

This approach based on comparable corpora considerably relieves the data acquisition bottleneck, but has the disadvantage that the results tend to lack accuracy in practice.

As an alternative, there is also the approach of identifying orthographically similar words (Koehn & Knight, 2002) which has the advantage that it does not even require a corpus. A simple word list will suffice. However, this approach works only for closely related languages, and has limited potential otherwise.

We propose here to generate dictionaries on the basis of foreign word occurrences in texts. As far as we know, this is a method which has not been tried before. When doing so, a single monolingual corpus can be used for all source languages for which it contains a sufficient number of foreign words. A constraint is that the target language must always be the language of the monolingual corpus,[2] which therefore all dictionaries have in common.

2 Approach and Language Resources

Starting from the observation that monolingual dictionaries typically include a considerable number of foreign words, the basic idea is to consider the most significant co-occurrences of a foreign word as potential translation candidates. This implies that the language of the underlying corpus must correspond to the target language, and that this corpus can be utilized for any source language for which word citations are well represented.

As the use of foreign language words in texts depends on many parameters, including writer, text type, status of language and cultural background, it is interesting to compare results when varying some of these parameters. However, due to the general scarceness of foreign word citations our approach requires very large corpora. For this reason, we were only able to vary two parameters, namely language and text type.

Some large enough corpora that we had at our disposal were the *Gigaword Corpora* from the Linguistic Data Consortium (Mendonça et al., 2009a; Mendonça et al., 2009b) and the *WaCky Corpora* described in Sharoff (2006), Baroni et al. (2009), and Ferraresi et al. (2010). From these, we selected the following for this study:

- French WaCky Corpus (8.2 GB)
- German WaCky Corpus (9.9 GB)
- Italian WaCky Corpus (10.4 GB)
- French Gigaword 2nd edition (5.0 GB)
- Spanish Gigaword 2nd edition (6.8 GB)

The memory requirements shown for each corpus relate to ANSI coded text only versions. We derived these from the original corpora by removing linguistic annotation (for the WaCky corpora) and XML markup, and by converting the coding from UTF8 to ANSI.

Both Gigaword corpora consist of newsticker texts from several press agencies. Newsticker text is a text type closely related to newspaper text. It is usually carefully edited, and the vocabulary is geared towards easy understanding for the intended readership. This implies that foreign word citations are kept to a minimum.

In contrast, the WaCky Corpora have been downloaded from the web and represent a great variety of text types and styles. Hence, not all texts can be expected to have been carefully edited, and mixes between languages are probably more frequent than with newsticker text.

As in this work English is the main source language, and as we have dealt with it as a target language already in Rapp & Zock (2010), we do not use the respective English versions of these corpora here. We also do not use the *Wikipedia XML Corpora* (Denoyer et al., 2006) as these greatly vary in size for different languages which makes comparisons across languages somewhat problematic. In contrast, the sizes of the above corpora are within the same order of magnitude (1 billion words each), which is why we do not control for corpus size here.

[2] Although in principle it would also be possible to determine relations between foreign words from different languages within a corpus, this seems not promising as the problem of data sparsity is likely to be prohibitive.

Concerning the number of foreign words within these corpora, we might expect that, given the status of English as the world's premiere language, English foreign words should be the most frequent ones in our corpora. As French and Spanish are also prominent languages, foreign words borrowed from them may be less frequent but should still be common, whereas borrowings from German and Italian are expected to be the least likely ones. From this point of view the quality of the results should vary accordingly. But of course there are many other aspects that are important, for example, relations between countries, cultural background, relatedness between languages, etc. As these are complex influences with intricate interactions, it is impossible to accurately anticipate the actual outcome. In other words, experimental work is needed. Let us therefore describe our approach.

For identifying word translations within a corpus, we assume that the strongest association to a foreign word is likely to be its translation. This can be justified by typical usage patterns of foreign words often involving, for example, an explanation right after their first occurrence in a text.

Associations between words can be computed in a straightforward manner by counting word co-occurrences followed by the application of an association measure on the co-occurrence counts. Co-occurrence counts are based on a text window comprising the 20 words on either side of a given foreign word. On the resulting counts we apply the log-likelihood ratio (Dunning, 1993). As explained by Dunning, this measure has the advantage to be applicable also on low counts, which is an important characteristic in our setting where the problem of data sparseness is particularly severe. This is also the reason why we chose a window size somewhat larger than the ones used in most other studies.

Despite its simplicity this procedure of computing associations to foreign words is well suited for identifying word translations. As mentioned above, we assume that the strongest association to a foreign word is its best translation.

We did this for words from five languages (English, French, German, Italian, and Spanish). The results are shown in the next section. In order to be able to quantitatively evaluate the quality of our results, we counted for all source words of a language the number of times the expected target word obtained the strongest association score.

Our expectations on what should count as a correct translation had been fixed before running the experiments by creating a gold standard for evaluation. We started from the list of 100 English words (nouns, adjectives and verbs) which had been introduced by Kent & Rosanoff (1910) in a psychological context.

We translated these English words into each of the four target languages, namely French, German, Italian, and Spanish. As we are at least to some extent familiar with these languages, and as the Kent/Rosanoff vocabulary is fairly straightforward, we did this manually. In cases where we were aware of ambiguities, we tried to come up with a translation relating to what we assumed to be the most frequent of a word's possible senses. In case of doubt we consulted a number of written bilingual dictionaries, the *dict.leo.org* dictionary website, and the translation services provided by Google and Yahoo. For each word, we always produced only a single translation. In an attempt to provide a common test set, the appendix shows the resulting list of *word equations* in full length for reference by interested researchers.

It should be noted that the concept of *word equations* is a simplification, as it does not take into account the fact that words tend to be ambiguous, and that ambiguities typically do not match across languages. Despite these shortcomings we nevertheless use this concept. Let us give some justification.

Word ambiguities are omnipresent in any language. For example, the English word *palm* has two meanings (*tree* and *hand*) which are usually expressed by different words in other languages. However, for our gold standard we must make a choice. We can not include two or more translations in one word equation as this would contradict the principle that all words in a word equation should share their main sense.

Another problem is that, unless we work with dictionaries derived from parallel corpora, it is difficult to estimate how common a translation is. But if we included less common translations in our list, we would have to give their matches a smaller weight during evaluation.

This, however, is difficult to accomplish accurately. This is why, despite their shortcomings, we use word equations in this work.

Evaluation of our results involves comparing a predicted translation to the corresponding word in the gold standard. We consider the predicted translation to be correct if there is a match, otherwise we consider it as false. While in principle possible, we do not make any finer distinctions concerning the quality of a match.

A problem that we face in our approach is what we call the *homograph trap*. What we mean by this term is that a foreign word occurring in a corpus of a particular language may also be a valid word in this language, yet possibly with a different meaning. For example, if the German word *rot* (meaning *red*) occurs in an English corpus, its occurrences can not easily be distinguished from occurrences of the English word *rot*, which is a verb describing the process of decay.

Having dealt with this problem in Rapp & Zock (2010) we will not elaborate on it here, rather we will suggest a workaround. The idea is to look only at a very restricted vocabulary, namely the words defined in our gold standard. There we have 100 words in each of the five languages, i.e. 500 words altogether. The question is how many of these words occur more often than once. Note, however, that apart from English (which was the starting point for the gold standard), repetitions can occur not only across languages but also within a language. For example, the Spanish word *sueño* means both *sleep* and *dream*, which are distinct entries in the list.

The following is a complete list of words showing either of these two types of repetitions, i.e. exact string matches (taking into account capitalization and accents): alto (4), bambino (2), Bible (2), bitter (2), casa (2), commando (2), corto (2), doux (2), duro (2), fruit (2), justice (2), lento (2), lion (2), long (2), luna (2), mano (2), memoria (2), mouton (2), religion (2), sacerdote (2), sueño (2), table (2), whisky (4).

However, as is obvious from this list, these repetitions are due to common vocabulary of the languages, with *whisky* being a typical example. They are not due to incidental string identity of completely different words. So the latter is not a problem (i.e. causing the identification of wrong translations) as long as we do not go beyond the vocabulary defined in our gold standard.

For this reason and because dealing with the full vocabulary of our (very large) corpora would be computationally expensive, we decided to replace in our corpora all words absent from the gold standard by a common designator for unknown words. Also, in our evaluations, for the target language vocabulary we only use the words occurring in the respective column of the gold standard.

So far, we always computed translations to single source words. However, if we assume, for example, that we already have word equations for four languages, and all we want is to compute the translations into a fifth language, then we can simply extend our approach to what we call the *product-of-ranks algorithm*. As suggested in Rapp & Zock (2010) this can be done by looking up the ranks of each of the four given words (i.e. the words occurring in a particular word equation) within the association vector of a translation candidate, and by multiplying these ranks. So for each candidate we obtain a product of ranks. We then assume that the candidate with the smallest product will be the best translation.[3]

Let us illustrate this by an example: If the given words are the variants of the word *nervous* in English, French, German, and Spanish, i.e. *nervous*, *nerveux*, *nervös*, and *nervioso*, and if we want to find out their translation into Italian, we would look at the association vectors of each word in our Italian target vocabulary. The association strengths in these vectors need to be inversely sorted, and in each of them we will look up the positions of our four given words. Then for each vector we compute the product of the four ranks, and finally sort the Italian vocabulary according to these products. We would then expect that the correct Italian translation, namely *nervoso*, ends up in the first position, i.e. has the smallest value for its product of ranks.

[3] Note that, especially in the frequent case of zero-co-occurrences, many words may have the same association strength, and rankings within such a group of words may be arbitrary within a wide range. To avoid such arbitrariness, it is advisable to assign all words within such a group the same rank, which is chosen to be the average rank within the group.

In the next section, we will show the results for this algorithm in addition to those for single source language words.

As a different matter, let us mention that for our above algorithm we do not need an explicit identification of what should count as a foreign word. We only need a list of words to be translated, and a list of target language words containing the translation candidates from which to choose. Overlapping vocabulary is permitted. If the overlapping words have the same meaning in both languages, then there is no problem and the identification of the correct translation is rather trivial as co-occurrences of a word with itself tend to be frequent. However, if the overlapping words have different meanings, then we have what we previously called a *homogaph trap*. In such (for small vocabularies very rare) cases, it would be helpful to be able to distinguish the occurrences of the foreign words from those of the homograph. However, this problem essentially boils down to a word sense disambiguation task (actually a hard case of it as the foreign word occurrences, and with them the respective senses, tend to be rare) which is beyond the scope of this paper.

3 Experimental Results and Evaluation

We applied the following procedure on each of the five corpora: The language of the respective corpus was considered the target language, and the vocabulary of the respective column in the gold standard was taken to be the target language vocabulary.

	Source Languages					
	DE	EN	FR	ES	IT	all
DE WaCky	–	54	22	18	20	48
ES Giga	9	42	37	–	29	56
FR Giga	15	45	–	20	14	49
FR WaCky	27	59	–	16	21	50
IT WaCky	17	53	29	27	–	56
Average	17.0	50.6	29.3	20.3	21.0	51.8

Table 1: Number of correctly predicted translations for various corpora and source languages. Column *all* refers to the parallel use of all four source languages using the product-of-ranks algorithm.

The other languages are referred to as the source languages, and the corresponding columns of the gold standard contain the respective vocabularies. Using the algorithm described in the previous section, for each source vocabulary the following procedure was conducted: For every source language word the target vocabulary was sorted according to the respective scores. The word obtaining the first rank was considered to be the predicted translation. This predicted translation was compared to the translation listed in the gold standard. If it matched, the prediction was counted as correct, otherwise as wrong.

Table 1 lists the number of correct predictions for each corpus and for each source language. These results lead us to the following three conclusions:

1) The noisier the better

We have only for one language (French) both a Gigaword and a WaCky corpus. The results based on the WaCky corpus are clearly better for all languages except Spanish. Alternatively, we can also look at the average performance for the five source languages among the three WaCky corpora, which is 30.3, and the analogous performance for the two Gigaword corpora, which is 26.4. These findings lend some support to our hypothesis that noisy web corpora are better suited for our purpose than carefully edited newsticker corpora, which are probably more successful in avoiding foreign language citations

2) English words are cited more often

In the bottom row, Table 1 shows for each of the five languages the scores averaged over all corpora. As hypothesized previously, we can take citation frequency as an indicator (among others) of the “importance” of a language. And citation frequency can be expected to correlate with our scores. With 50.6, the average score for English is far better than for any other language, thereby underlining its special status among world languages. With an average score of 29.3 French comes next which confirms the hypothesis that it is another world language receiving considerable attention elsewhere. Somewhat surprising is the finding that Spanish can not keep up with French and obtains an average score of 20.3 which is even lower than the 21.0 for Italian. A possible explanation is the fact that we are only dealing with European languages here, and that the cultural influence of the Roman Empire and Italy has been so considerable in Europe that it may well account for this. So the status of Spanish in the world may not be well reflected in our selection of corpora. Finally, the average score of 17.0 for German shows that it is the least cited language in our selection of languages. Bear in mind, though, that German is the only clearly Germanic language here, and that its vocabulary is very different from that of the other languages. These are mostly Romanic in type, with English somewhere in between. Therefore, the little overlap in vocabulary might make it hard for French, Italian, and Spanish writers to understand and use German foreign words.

3) Little improvement for several source words

The right column in Table 1 shows the scores if (using the product-of-ranks algorithm) four source languages are taken into account in parallel. As can be seen, with an average score of 51.8 the improvement over the English only variant (50.6) is minimal. This contrasts with the findings described in Rapp & Zock (2010) where significant improvements could be achieved by increasing the number of source languages. So this casts some doubt on these. However, as English was not considered as a source language there, the performance levels were mostly between 10 and 20, leaving much room for improvement. This is not the case here, where we try to improve on a score of around 50 for English. Remember that this is a somewhat conservative score as we count correct but alternative translations, as errors. As this is already a performance much closer to the optimum, making further performance gains is more difficult. Therefore, perhaps we should take it as a success that the product-of-ranks algorithm could achieve a minimal performance gain despite the fact that the influence of the non-English languages was probably mostly detrimental.

Having analyzed the quantitative results, to give a better impression of the strengths and weaknesses of our algorithm, for the (according to Table 1) best performing combination of corpus and language pair, namely the French WaCky corpus, English as the source language and French as the target language, Table 2 shows some actual source words and their computed translations.

ESW	CF	ET	RE	CT
cabbage	9	chou	1	chou
blossom	25	fleur	73	commande
carpet	39	tapis	1	tapis
bitter	59	amer	1	amer
hammer	67	marteau	1	marteau
bread	82	pain	1	pain
citizen	115	citoyen	1	citoyen
bath	178	bain	1	bain
butterfly	201	papillon	1	papillon
eat	208	manger	1	manger
butter	220	beurre	59	terre
eagle	282	aigle	1	aigle
cheese	527	fromage	1	fromage
cold	539	froid	1	froid
deep	585	profond	1	profond
cottage	624	cabanon	1	cabanon
earth	702	terre	53	tabac
child	735	enfant	1	enfant
bed	806	lit	2	table
beautiful	923	beau	1	beau
care	1267	soin	1	soin
hand	1810	main	2	main
city	2610	ville	1	ville
girl	2673	fille	1	fille
green	2861	vert	1	vert
blue	2914	bleu	1	bleu
hard	3615	dur	1	dur
black	9626	noir	1	noir
Bible	17791	Bible	1	Bible
foot	23548	pied	8	siffler
chair	24027	chaise	1	chaise
fruit	38544	fruit	1	fruit

Table 2: Results for the language pair English → French. The meaning of the columns is as follows: ESW = English source word; CF = corpus frequency of English source word; ET = expected translation according to gold standard; RE = computed rank of expected translation; CT = computed translation.

4 Summary and Future Work

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

Whereas Rapp & Zock (2010) dealt only with an English corpus, the current work shows that this methodology is applicable to a wide range of languages and corpora. We were able to shed some light on criteria influencing performance, such as the selection of text type and the direction of a language pair. For example, it is more promising to look at occurrences of English words in a German corpus rather than the other way around. Because of the special status of English it is also advisable to use it as a pivot wherever possible.

Perhaps surprisingly, the work may have implications regarding cognitive models of second language acquisition. The reason is that it describes how to acquire the vocabulary of a new language from a mixed corpus. This is relevant as traditional foreign language teaching (involving explanations in the native tongue and vocabulary learning using bilingual word lists) can be considered as providing such a mixed corpus.

Regarding future work, let us outline a plan for the construction of a universal dictionary of all languages which are well enough represented on the web.[4] There might be some chance for it, because the algorithm can be extended to work with standard search engines and is also suitable for a bootstrapping approach.

Let us start by assuming that we have a large matrix where the rows correspond to the union of the vocabularies of a considerable number of languages, and the columns correspond to these languages themselves. We presuppose no prior translation knowledge, so that the matrix is completely empty at the beginning (although prior knowledge could be useful for the iterative algorithm to converge).

STEP 1: For each word in the vocabulary we perform a search via a search engine such as Google, preferably in an automated fashion via an application programming interface (API). Next, we retrieve as many documents as possible, and separate them according to language.[5] Then, for each language for which we have obtained the critical mass of documents, we apply our algorithm and compute the respective translations. These are entered into the matrix. As we are interested in word equations, we assume that translations are symmetric. This means that each translation identified can be entered at two positions in the matrix. So at the end of step 1 we have for each word the translations into a number of other languages, but this number may still be small at this stage.

STEP 2: We now look at each row of the matrix and feed the words found within the same row into the product-of-ranks algorithm. We do not have to repeat the Google search, as step 1 already provided all documents needed. Because when looking at several source words we have a better chance to find occurrences in our documents, this should give us translations for some more languages in the same row. But we also need to recompute the translations resulting from the previous step as some of them will be erroneous e.g. for reasons of data sparseness or due to the homograph trap.

STEP 3: Repeat step 2 until as many matrix cells as possible are filled with translations. We hope that with each iteration completeness and correctness improve, and that the process converges in such a way that the (multilingual) words in each row disambiguate each other, so that ultimately each row corresponds to an unambiguous concept.

Acknowledgments

Part of this research was supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme. We thank the WaCky group for making available their excellent Web corpora, and the Linguistic Data Consortium for providing the Gigaword Corpora. We also wish to thank Lourdes Callau, Maria Dolores Jimenez Lopez, and Lilica Voicu for their support in acquiring the LDC corpora.

[4] Note that this plan could also be adapted to other methodologies (such as Rapp, 1999), and may be more promising with these.

[5] If the language identification markup within the retrieved documents turns out to be unreliable (which is unfortunately often the case in practice), standard language identification software can be used.

References

Armstrong, Susan; Kempen, Masja; McKelvie, David; Petitpierre, Dominique; Rapp, Reinhard; Thompson, Henry (1998). Multilingual Corpora for Cooperation. *Proceedings of the 1st International Conference on Linguistic Resources and Evaluation (LREC), Granada,* Vol. 2, 975–980.

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.

Brown, Peter; Cocke, John; Della Pietra, Stephen A.; Della Pietra, Vincent J.; Jelinek, Frederick; Lafferty, John D.; Mercer, Robert L.; Rossin, Paul S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2), 79–85.

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.

Denoyer, Ludovic; Gallinari, Pattrick (2006). The Wikipedia XML Corpus. *SIGIR Forum*, 40(1), 64–69.

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

Ferraresi, Adriano; Bernardini, Silvia; Picci, Giovanni; Baroni, Marco (2010). Web corpora for bilingual lexicography: a pilot study of English/ French collocation extraction and translation. In Xiao, Richard (ed.): *Using Corpora in Contrastive and Translation Studies*. Newcastle: Cambridge Scholars Publishing.

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

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

Harris, Zelig S. (1954). Distributional structure. *WORD*, 10:146–162.

Kent, Grace Helen; Rosanoff , A.J. (1910). A study of association in insanity. American Journal of Insanity 67:317–390.

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

Koehn, Philipp; Hoang, Hieu; Birch, Alexandra; Callison-Burch, Chris; Federico, Marcello; Bertoldi, Nicola; Cowan, Brooke; Shen, Wade; Moran, Christine; Zens, Richard; Dyer, Chris; Bojar, Ondřej; Constantin, Alexandra; Herbst, Evan (2007). Moses: Open source toolkit for statistical machine translation. In: *Proceedings of ACL*, Prague, demonstration session, 177–180.

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

Mendonça, Angelo, Graff, David, DiPersio, Denise (2009a). *French Gigaword Second Edition.* Linguistic Data Consortium, Philadelphia.

Mendonça, Angelo, Graff, David, DiPersio, Denise (2009b). *Spanish Gigaword Second Edition.* Linguistic Data Consortium, Philadelphia.

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

Och, Franz Josef; Ney, Hermann (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1), 19–51.

Pantel, Patrick; Lin, Dekang (2002). Discovering word senses from text. In: *Proceedings of ACM SIGKDD*, Edmonton, 613–619

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; Martín Vide, Carlos (2007). Statistical machine translation without parallel corpora. In: Georg Rehm, Andreas Witt, Lothar Lemnitzer (eds.): *Data Structures for Linguistic Resources and Applications. Proceedings of the Biennial GLDV Conference 2007*. Tübingen: Gunter Narr Verlag. 231–240.

Rapp, Reinhard; Zock, Michael (2010). Utilizing Citations of Foreign Words in Corpus-Based Dictionary Generation. *Proceedings of NLPIX 2010.*

Sharoff, Serge (2006). Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini (eds.): WaCky! *Working papers on the Web as Corpus*. Gedit, Bologna, http://wackybook.sslmit.unibo.it/

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 5th International Conference on Language Resources and Evaluation* ***(LREC 2006).*** 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.

Appendix: Gold Standard of 100 Word Equations

	ENGLISH	GERMAN	FRENCH	SPANISH	ITALIAN
1	anger	Wut	colère	furia	rabbia
2	baby	Baby	bébé	bebé	bambino
3	bath	Bad	bain	baño	bagno
4	beautiful	schön	beau	hermoso	bello
5	bed	Bett	lit	cama	letto
6	Bible	Bibel	Bible	Biblia	Bibbia
7	bitter	bitter	amer	amargo	amaro
8	black	schwarz	noir	negro	nero
9	blossom	Blüte	fleur	flor	fiore
10	blue	blau	bleu	azul	blu
11	boy	Junge	garçon	chico	ragazzo
12	bread	Brot	pain	pan	pane
13	butter	Butter	beurre	mantequilla	burro
14	butterfly	Schmetterling	papillon	mariposa	farfalla
15	cabbage	Kohl	chou	col	cavolo
16	care	Pflege	soin	cuidado	cura
17	carpet	Teppich	tapis	alfombra	tappeto
18	chair	Stuhl	chaise	silla	sedia
19	cheese	Käse	fromage	queso	formaggio
20	child	Kind	enfant	niño	bambino
21	citizen	Bürger	citoyen	ciudadano	cittadino
22	city	Stadt	ville	ciudad	città
23	cold	kalt	froid	frío	freddo
24	command	Kommando	commande	comando	comando
25	convenience	Bequemlichkeit	commodité	conveniencia	convenienza
26	cottage	Häuschen	cabanon	casita	casetta
27	dark	dunkel	foncé	oscuro	buio
28	deep	tief	profond	profundo	profondo
29	doctor	Arzt	médecin	médico	medico
30	dream	Traum	rêve	sueño	sogno
31	eagle	Adler	aigle	águila	aquila
32	earth	Erde	terre	tierra	terra
33	eat	essen	manger	comer	mangiare
34	foot	Fuß	pied	pie	piede
35	fruit	Frucht	fruit	fruta	frutta
36	girl	Mädchen	fille	chica	ragazza
37	green	grün	vert	verde	verde
38	hammer	Hammer	marteau	martillo	martello
39	hand	Hand	main	mano	mano
40	handle	Griff	poignée	manejar	maniglia
41	hard	hart	dur	duro	duro
42	head	Kopf	tête	cabeza	testa
43	health	Gesundheit	santé	salud	salute
44	heavy	schwer	lourd	pesado	pesante

45	high	hoch	élevé	alto	alto
46	house	Haus	maison	casa	casa
47	hungry	hungrig	affamé	hambriento	affamato
48	joy	Freude	joie	alegría	gioia
49	justice	Gerechtigkeit	justice	justicia	giustizia
50	King	König	roi	rey	re
51	lamp	Lampe	lampe	lámpara	lampada
52	light	Licht	lumière	luz	luce
53	lion	Löwe	lion	león	leone
54	long	lang	long	largo	lungo
55	loud	laut	fort	alto	alto
56	man	Mann	homme	hombre	uomo
57	memory	Gedächtnis	mémoire	memoria	memoria
58	moon	Mond	lune	luna	luna
59	mountain	Berg	montagne	montaña	montagna
60	music	Musik	musique	música	musica
61	mutton	Hammel	mouton	cordero	montone
62	needle	Nadel	aiguille	aguja	ago
63	nervous	nervös	nerveux	nervioso	nervoso
64	ocean	Ozean	océan	océano	oceano
65	oven	Backofen	four	horno	forno
66	priest	Priester	prêtre	sacerdote	sacerdote
67	quick	schnell	rapide	rápido	rapido
68	quiet	still	tranquille	tranquilo	tranquillo
69	red	rot	rouge	rojo	rosso
70	religion	Religion	religion	religión	religione
71	river	Fluss	rivière	río	fiume
72	rough	rau	rugueux	áspero	ruvido
73	salt	Salz	sel	sal	sale
74	scissors	Schere	ciseaux	tijeras	forbici
75	sheep	Schaf	mouton	oveja	pecora
76	short	kurz	courte	corto	corto
77	sickness	Krankheit	maladie	enfermedad	malattia
78	sleep	schlafen	sommeil	sueño	dormire
79	slow	langsam	lent	lento	lento
80	smooth	glatt	lisse	liso	liscio
81	soft	weich	doux	suave	morbido
82	soldier	Soldat	soldat	soldado	soldato
83	sour	sauer	acide	agrio	acido
84	spider	Spinne	araignée	araña	ragno
85	square	Quadrat	carré	cuadrado	quadrato
86	stomach	Magen	estomac	estómago	stomaco
87	street	Straße	rue	calle	strada
88	sweet	süß	doux	dulce	dolce
89	table	Tisch	table	mesa	tavolo
90	thief	Dieb	voleur	ladrón	ladro
91	thirsty	durstig	soif	sediento	assetato
92	tobacco	Tabak	tabac	tabaco	tabacco
93	whisky	Whisky	whisky	whisky	whisky
94	whistle	pfeifen	siffler	silbar	fischiare
95	white	weiß	blanc	blanco	bianco
96	window	Fenster	fenêtre	ventana	finestra
97	wish	Wunsch	désir	deseo	desiderio
98	woman	Frau	femme	mujer	donna
99	work	arbeiten	travail	trabajo	lavoro
100	yellow	gelb	jaune	amarillo	giallo