Identifying Word Translations from Comparable Documents Without a Seed Lexicon

Reinhard Rapp, Serge Sharoff, Bogdan Babych
Centre for Translation Studies, University of Leeds
Leeds, LS2 9JT, United Kingdom
[R.Rapp, S.Sharoff, B.Babych,]@leeds.ac.uk

Abstract

The extraction of dictionaries from parallel text corpora is an established technique. However, as parallel corpora are a scarce resource, in recent years the extraction of dictionaries using comparable corpora has obtained increasing attention. In order to find a mapping between languages, almost all approaches suggested in the literature rely on a seed lexicon. The work described here achieves competitive results without requiring such a seed lexicon. Instead it presupposes mappings between comparable documents in different languages. For some common types of textual resources (e.g. encyclopedias or newspaper texts) such mappings are either readily available or can be established relatively easily. The current work is based on Wikipedias where the mappings between languages are determined by the authors of the articles. We describe a neural-network inspired algorithm which first characterizes each Wikipedia article by a number of keywords, and then considers the identification of word translations as a variant of word alignment in a noisy environment. We present results and evaluations for eight language pairs involving Germanic, Romanic, and Slavic languages as well as Chinese.

Keywords: term alignment, identification of word translations, dictionary extraction, comparable corpora

1. Introduction

In a globalized world with easy web access to documents in a multitude of languages, comprehensive information searches require translation capabilities for a large number of language pairs. Essential prerequisites for translation are bilingual dictionaries. Apart from traditional lexicographic methods, they can be automatically generated from parallel corpora by applying algorithms for sentence and word alignment. But especially for language pairs involving lesser-used languages parallel corpora are still a scarce resource, and attempts to mine parallel text segments from pairs of monolingual corpora could only slightly relieve the problem (Munteanu & Marcu, 2005).

Therefore, to resolve the data acquisition bottleneck, researchers came up with the idea of generating dictionaries directly from comparable corpora. Comparable corpora are far more common than parallel corpora, and are for many languages available in large quantities e.g. in the form of newspaper texts or encyclopedias such as Wikipedia, which is available for about 280 languages[1] But not only are comparable corpora easier to acquire, we also need fewer of them: Whereas in the case of comparable corpora (if they are all in the same domain) usually one corpus per language suffices, for parallel corpora typically one corpus per language pair is required (unless translations of the same corpus are available for many languages). This means that instead of a linear increase there is a quadratic increase with the number of languages, which explains why methods based on comparable corpora have the potential to significantly diminish the data acquisition bottleneck.

The basic assumption underlying most approaches based on comparable corpora is that across languages there is a correlation between the co-occurrence patterns of words which are translations of each other. If, for example, in language *A* two words co-occur more often than expected by chance, then their translated equivalents in language *B* should also co-occur more frequently than expected. The validity of this co-occurrence constraint is obvious for parallel corpora, but it also holds for non-parallel corpora. It can be expected that this constraint will work best with parallel corpora, second-best with comparable corpora, and somewhat worse with unrelated corpora. Robustness is not a major issue in any of these cases. In contrast, when applying sentence alignment algorithms to parallel corpora, omissions, insertions, and transpositions of text segments can have critical negative effects. However, the co-occurrence constraint when applied to comparable corpora is much weaker than the word-order constraint as used with parallel corpora. This is why it is much more difficult to come up with reasonable results.

The current paper can be seen as a continuation of our previous work (Rapp, 1999). Due to space constraints, for an overview on other related work let us point to Laws et al. (2010). However, apart from comparatively limited methods which are based on cognates and therefore only work for closely related languages, almost all previous approaches have in common that they presuppose an initial dictionary (bilingual lexicon of seed words) in order to be able to relate between languages. In contrast, the approach which we present here does not require such a lexicon, but instead assumes the availability of aligned comparable documents. For some common text types this is not an unreasonable requirement. For example, in Wikipedia articles of different languages are aligned via the so-called interlanguage links, and newspaper articles can be aligned via their dates of publication in combination with some basic topic detection software.

[1] http://meta.wikimedia.org/wiki/List_of_Wikipedias

2. Approach

Like most previous ones our approach is also based on the assumption that there is a correlation between the patterns of word-co-occurrence across languages. However, instead of presupposing a bilingual dictionary it only requires pre-aligned comparable documents, i.e. small or medium sized documents across languages which are known to deal with similar topics. This can be, for example, newspaper articles, scientific papers, contributions to discussion groups, or encyclopaedic articles. As Wikipedia is a large resource and readily available for many languages, and to be able to compare our results to recent related work which also uses Wikipedia (Laws et al., 2010), we decided to base our study on this encyclopaedia. Our algorithm is (apart from word segmentation issues) largely language independent and should lead to similarly good results for any languages where Wikipedias of reasonable size are available. Some statistics of the Wikipedias used in this paper are shown in Table 1 (Wikipedia dumps were downloaded from http://dumps.wikimedia.org/ between November 4 and 12, 2011).

LANGUAGE	ID	MILLION TOKENS	ARTICLES	EN IWIKI LINKS
Chinese	zh	101	137179	87389
Dutch	nl	163	435716	290979
English	en	1440	2524134	n/a
French	fr	459	838771	541715
German	de	563	1114696	603437
Portuguese	pt	156	361204	245102
Russian	ru	268	609525	345195
Spanish	es	365	664097	438864
Ukrainian	uk	81	214403	139827

Table 1: Wikipedia statistics.

The Wikipedias have the so-called *interlanguage links* which connect two articles in different languages. Therefore, if a headword is dealt with in several languages, a special interlanguage tag (iwiki) is usually placed in the respective article. For example, the English article on the headword *Depth-of-field adapter*[2] contains iwiki links such as:

Czech: DOF adaptér
German: 35-Millimeter-Adapter
Korean: DOF 어댑터
Japanese: DOF アダプタ
Russian: DOF-адаптер

[2] This is an image converter allowing the exchange of camera lenses, thereby providing a shallow depth of field.

The right column in Table 1 refers to the number of iwiki links from the various languages into English.
Given that each Wikipedia community contributes in their own language, only in rare cases an article connected in this way will be a simple translation of the English article, and in most cases the contents will be rather different. On the positive side, the link structure of the interlanguage links tends to be quite dense, see the above table. It should be mentioned that the set of headwords connected by these links can already be considered as a raw dictionary of mainly nouns and proper nouns, which in principle could be used for evaluation purposes. However, in this work we decided to use evaluation data from an independent source.

2.1 Preprocessing steps

After download, each Wikipedia was minimally processed to extract the plain text contents of the articles. In this process all templates, e.g. `infoboxes', as well as tables were removed, and we kept only the webpages with more than 500 characters of running text (including white space). We maintained the iwiki links to the English webpages as well as `Categories', though the latter were not used in the process discussed below.
Linguistic processing steps included tokenisation, tagging and lemmatisation using the default UTF-8 versions of the respective TreeTagger resources (Schmid, 1994) for all languages except Russian and Ukrainian, for which tagging and lemmatisation was done using our own tools (Sharoff, et al, 2008) based on TnT (Brants, 2000) and CST lemmatiser (Jongejan & Dalianis, 2009). Given that the tokeniser for Chinese used in TreeTagger (which is in turn our own development) uses the simplified script, the contents of the Chinese Wikipedia was converted to the simplified characters for uniformity reasons.

2.2 Alignment steps

As these documents are typically not translations of each other, we cannot apply the usual procedure and tools as available for parallel texts (e.g. the Gale & Church sentence aligner and the Giza++ word alignment tool). Instead we conduct a two step procedure:

1) We first extract salient terms from each of the documents.

2) We then align these terms across languages using an approach inspired by a connectionist (Rumelhart & McClelland, 1987) WINner-Takes-It-All Network (WINTIAN algorithm).

The procedure for term extraction is based on using the frequency list of the entire Wikipedia to measure the keyness of words in each individual article. The articles are usually short, with an average length of about 500 words, so we use the log-likelihood score as a measure of keyness, since it has been shown to be robust to small numbers of instances (Rayson & Garside, 2000). For example, the keywords extracted for English and German for the above-mentioned article (Depth-of-field adapter) are shown in Table 2.

English		
LL-score	N	Term
288.73	21	adapter
173.00	17	lens
151.71	10	camcorder
137.11	17	screen
120.45	18	focus
94.43	9	flip
83.97	11	camera
80.00	15	image
58.59	5	macro
50.47	4	35mm
38.83	2	plano-convex
34.87	3	translucent
33.44	2	vignetting
31.12	5	mount
25.31	3	photographic
25.18	3	texture
22.68	2	flange
21.84	2	aberration
21.65	2	post-production
20.27	2	chromatic
20.24	3	module
20.01	2	upside
19.88	3	mirror
19.52	2	zoom
19.03	2	prism
18.89	3	monitor
18.41	2	blur
16.80	5	must
16.80	3	correct
15.35	2	Canon
15.35	3	frame
15.33	3	attach

German		
LL-score	N	Term
253.75	14	Mattscheibe
116.09	5	35-mm-Adapter
46.58	3	Körnung
43.61	2	35-Millimeter-Adapter
38.84	3	Adapter
37.72	2	HD-Auflösung
32.35	8	Bild
31.65	3	Linse
29.86	3	Objektiv
29.44	4	Kamera
28.01	3	statisch
27.76	2	Schärfentiefe
25.72	2	Videokamera
24.30	2	Spiegelreflexkamera
23.54	2	Sucher
17.84	2	bewegt
17.26	3	Hersteller
16.94	2	Hundert
16.10	2	Scheibe
15.25	2	Einschränkung

Table 2: English and German keywords for the Wikipedia article 'Depth-of-field adapter'. (LL = log-likelihood; N = term frequency in document.)

According to Rayson & Garside (2000) the threshold of 15.13 for the log-likelihood score is a conservative recommendation for statistical significance.

The WINTIAN algorithm is used for establishing term alignments across languages. As a detailed technical description is given in Rapp (1996: 108), we only briefly describe this algorithm here, thereby focusing on the neural network analogy. The algorithm can be considered as an artificial neural network where the nodes are all English and German words occurring in the keyword lists. Each English word has connections to all German words whose weights are all one at the beginning, but will be a measure of the translation probabilities after the completion of the algorithm. One after the other, the network is fed with the pairs of corresponding keyword lists. Each German word activates the corresponding German node with an activity of one. This activity is then propagated to all English words occurring in the corresponding list of keywords. The distribution of the activity is not equal, but in proportion to the connecting weights. This unequal distribution has no effect at the beginning when all weights are one, but later on leads to rapid activity increases for pairs of words which often occur in corresponding keyword lists. Of course it is assumed that these are translations of each other. Using Hebbian learning (Rumelhart & McClelland, 1987) the activity changes are stored in the connections. We use a heuristic to avoid the effect that frequent keywords dominate the network: When more than 50 of the connections to a particular English node have weights higher than one, the weakest 20 of them are reset to one. This way only translations which are frequently confirmed can build up high weights

Let us look at an example. Assume we have the (very short) English keyword list '*bank money*' corresponding to the German list '*Bank Geld*', and another English list '*bank river*' corresponding to the German '*Bank Fluss*'. When in the first cycle the network receives the first pair of keywords, it cannot decide whether '*bank*' corresponds to '*Bank*' or to '*Geld*', so will assign each possibility an activity of 0.5. So both weights will be increased equally. But in the second cycle when it comes to distributing the activity of '*bank*', the weight to '*Bank*' will be stronger than the one to '*Fluss*'. Therefore '*Bank*' will receive more activity, and the respective weight will become even stronger, in effect correctly disambiguating the ambiguous English word.

It turned out that the algorithm shows a robust behaviour in practice, which is important as the corresponding keyword lists are very noisy and may well contain less than 20% words which are actually translations of each other. Reasons are that corresponding articles are often written from different perspectives and can considerably vary in length. (To give an example, the descriptions of politicians tend to be very country specific). Nevertheless the algorithm is capable of grasping the regularities, and often comes up with reasonable results.

2.3 Vocabularies

The WINTIAN algorithm needs as input vocabularies of the source and the target language. For each language, we constructed these as follows: Based on the keyword lists for the respective Wikipedia, we counted the number of occurrences of each keyword, and then applied a threshold of five, i.e. all keywords with a lower frequency were eliminated. The reasoning behind this is that rare keywords are of not much use due to data sparseness.[3] To this vocabulary we added all words of the applicable gold standard(s) relating to the respective language (i.e. including the Google translations, and, if applicable for a language, their manual corrections, and the TS100 test set). Note that adding the words from the gold standard(s) means only a modest increase in vocabulary size as most of them easily meet the frequency threshold. Applying this procedure led to the vocabulary sizes as shown in Table 3.

LANGUAGE	ID	MILLION TOKENS	VOCABULARY SIZE
Chinese	zh	101	36623
Dutch	nl	163	58563
English	en	1440	133806
French	fr	459	101399
German	de	563	144251
Portuguese	pt	156	50003
Russian	ru	268	80940
Spanish	es	365	89732
Ukrainian	uk	81	30888

Table 3: Corpus and vocabulary sizes.

The vocabularies for larger Wikipedias are more comprehensive because more keywords meet the minimum frequency. As the gold standard words are included in any case, the selection task for the WINTIAN algorithm is somewhat easier for languages with a smaller Wikipedia, as the choice of words is more limited. Although at the above vocabulary sizes this is hardly noticeable, it would be an important factor for very small vocabularies. As a consequence, not only corpus size but also vocabulary size is of importance when comparing different algorithms, a fact which is sometimes overlooked.

[3] In corpus based studies sometimes thresholds of e.g. 50 are recommended. However, as here we consider keywords which have a higher information content than an average token in a corpus, it makes sense to use a lower threshold.

3. Evaluation setup

Our aim was to have a gold standard of word equations to test the predicted translation equivalents as computed by the WINTIAN algorithm. The source language words in the gold standard were supposed to be systematically derived from a large corpus, covering a wide range of frequencies, parts of speech, and variances of their distribution. In addition, the corpus from which the gold standard was derived was supposed to be completely separate from the development set (Wikipedia). The limitation of this method is, however, that translations were generated by *Google Translate*, and then manually checked, and only one of several possible translations of the English words is included into the gold standard.

3.1 Preparing the gold standards

For a quantitative evaluation we used two data sets consisting of word equations. The first gold standard is the TS100 test set as described in Laws et al. (2010) and previously used by Rapp (1999) for the German-English pair. It comprises 100 English words together with their German translations.

As the TS100 test set is rather small, we developed a larger test set comprising 1000 items. We began with a list of words extracted from the British National Corpus (BNC) by Adam Kilgarriff for the purpose of examining distributional variability. This list is described at http://kilgarriff.co.uk/bnc-readme.html. It contains 8187 words which are all those that occur at least 100 times in a 10.1-million word subset of the BNC, comprising those documents which are at least 5000 words in length. Kilgarriff's main idea was to look at variation in frequencies across 2018 5000-word segments. Thus the items give us data about frequency and variability for future experiments, although at present we have not used this information.

Since these items are words, not lemmas, the next step was to pick uninflected forms by using the CLAWS tags attached. Taking the tagtypes shown in Table 4 and for each multi-tagged word keeping only the highest in the list (most frequent) gives a total of 3857 entries. (We excluded items that don't begin with a letter, and multi-word units with underscore or hyphen as delimiter.) From these we selected 1001 at random. (One item, "q.v.", was dropped as unsuitable, leaving a round thousand.) Numbers of items in each postag category are shown in Table 4.

The resulting list of 1000 English words was translated to the eight other languages (see Table 3) using *Google Translate*. For three of the languages, namely German, Russian, and Ukrainian, these translations were corrected by native speakers. The number of items which needed correction turned out to be in the order of 100 per language. The translations for all other languages remained uncorrected.

PART OF SPEECH	NUMBER
aj0 Adjective	237
av0 Adverb	93
crd Cardinal number	12
nn0 Collective (or mass) noun	15
nn1 Singular noun	546
ord Ordinal number	3
prp Preposition	10
vbi Verb "be" infinitive	1
vdi Verb "do" infinitive	0
vhi Verb "have" infinitive	1
vvb Verb base-form	7
vvi Verb infinitive	75

Table 4: Occurrences of postag categories.

4. Results and evaluation

Using the WINTIAN algorithm, the English translations for all 144,251 words occurring in the German vocabulary have been computed. Table 5 shows sample results for three German words.

GIVEN GERMAN WORD	STRASSE	
EXPECTED TRANSLATION	STREET	
	LL-SCORE	TRANSLATION
1	215.3	road
2	148.2	street
3	66.0	traffic
4	46.0	Road
5	42.6	route
6	34.6	building

GIVEN GERMAN WORD	KRANKHEIT	
EXPECTED TRANSLATION	SICKNESS	
	LL-SCORE	TRANSLATION
1	236.4	disease
2	105.3	symptom
3	61.6	illness
4	50.8	epidemic
5	44.0	treatment
6	39.1	genetic

GIVEN GERMAN WORD	GELB	
EXPECTED TRANSLATION	YELLOW	
	LL-SCORE	TRANSLATION
1	200.7	yellow
2	89.5	Yellow
3	17.9	green
4	13.8	tree
5	13.4	bright
6	13.1	pigment

Table 5: Sample results.

4.1 Comparison with other work

For a quantitative evaluation, we verified in how many cases our algorithm had assigned the expected translation (as provided by the gold standard) the top rank among all 133,806 translation candidates. (Candidates are all words occurring in the English vocabulary, see section 2.4.)
Table 6 compares our results to those of Laws et al. (2010) which represent the current state of the art, and to the Rapp (1999) baseline.[4] (All results are based on the English and German Wikipedia corpora.)

SYSTEM	ACCURACY
Baseline (Rapp, 1999)	50%
State of the art (Laws et al., 2010)	52%
Current approach	61%

Table 6: Comparison of systems.

As can be seen, the new approach outperforms the previous ones, although it should be noted that the Wikipedia contents have changed over time and that a comparison based on only 100 test words can only give a rough indication.[5]
A problem with our approach is that some words of the source language (typically ones with unspecific meaning) never make it for keywords, so no translations can be computed for them. In the case of the TS100 test set, this was the case for 7 out of 100 source language (i.e. German) words, that is the WINTIAN algorithm only had a chance to come up with the correct result in 93 cases. (But the above accuracy of 61% of course refers to all 100 test items.)
To reduce this problem, we experimented with setting the log-likelihood threshold for keywords lower, which, however, reduced the specificity of the keywords and consequently led to a lower overall accuracy (e.g. in the order of 40% for a threshold of zero).[6]
Let us mention that the results in Table 6 refer to exact matches with the word equations in the gold standard. As in reality due to word ambiguity other translations might also be acceptable (e.g. for 'Straße' not only 'street' but also 'road' would be acceptable, see Table 5), these figures are conservative and can be seen as a lower bound of the actual performance.
Another reason why the figures are conservative is translation asymmetry: To be comparable between languages our gold standard started with a list of English words, which were translated into the other language. However,

[4] Note that the scores reported in Rapp (1999) were based on different corpora and a proprietary seed lexicon which is why this work had been replicated by Laws et al. (2010) using Wikipedia and a freely available lexicon.
[5] We could not easily compare with the TS1000 testset provided by Laws et al. (2010) as this adds some more sophistication (parts of speech and multiple translations) to the evaluation process, whereas we, as we are dealing with many language, wanted to keep the evaluation process simple.
[6] Variable thresholds depending on word frequency might reduce the problem but this has not been implemented.

in this paper we are considering the translation directions from the foreign languages into English (reverse direction to be covered in future work). However, if the most common translation of source language word A is target language word B, then, due to asymmetry, the most common backtranslation of B is not necessarily A. This means our gold standard is suboptimal when used in the direction from the foreign language to the source language.

Concerning our results, it may also be of interest in how many cases the expected translation was not ranked first, but ended up on other positions of the computed lists (as exemplified in Table 5). For the TS100 test set, rank 2 was obtained in nine cases, and rank 3 in one case. Ranks 4 to 10 were obtained in no case.

4.2 Application to other languages

In comparison to Laws et al. (2010) our approach is knowledge-poor which means that, apart from word segmentation and lemmatization (which improves results but is not essential) it does not require any linguistic processing. It also does not require a lexicon of seed words (typically comprising at least 10,000 words). For these reasons and because Wikipedia provides document alignments for many languages, it was straightforward to apply our algorithm to a number of other languages.

However, for accurate measurements a gold standard larger than the TS100 test set was desirable, and this had to be extended to the new languages, as described in section 3.1. Applying our algorithm to the language pairs German → English, Russian → English, and Ukrainian → English and comparing the outcome with the manually corrected versions of the gold standard led to the results as shown in Table 7.

	DE→ EN	RU→ EN	UK→ EN
KW	925	873	817
1	381	331	229
2	43	42	25
3	12	11	11
4	5	8	9
5	8	5	2
6	2	3	5
7	4	1	5
8	1	2	1
9	0	1	1
10	0	0	1

Table 7: Results for three language pairs where the gold standard had been verified by native speakers.

Here in the second row 'KW' means the number of source language words in the gold standard (i.e. out of 1000) which in the keyword list of the corresponding source language Wikipedia actually occurred (see section 4.1), i.e. where the WINTIAN algorithm had a chance to compute English translations. The numbers in column 1 are ranks, and the figures in the other columns indicate the number of expected translations which ended up on the respective rank. For example, for the language pair German to English, 381 of the altogether 1000 expected translations (as taken from the gold standard) ended up on rank 1, 43 on rank 2, and so on. The accuracy for German is 38.1% as 381 of 1000 items were predicted correctly. This is considerably lower than our result for the TS100 test set where we had an accuracy of 61%.

Note, however, that this drop in accuracy for the larger test set is in line with expectations. The TS100 test set contains almost only very common words which have a high corpus frequency and are thus easy to predict. In contrast, by its construction the 1000 item test set (random selection from Adam Kilgarriff's large word list) represents a much wider frequency spectrum. Laws et al. (2010) made a similar observation (i.e. drop in accuracy) with their larger test set, although theirs consists of the top 1000 most frequent Wikipedia words only and should therefore be easier to deal with than ours.

If we now compare the results for the three language pairs, as expected we can observe an improvement in accuracy with an increase in the size of the respective version of Wikipedia (see Table 1). On the other hand, there are numerous other influences, including the relatedness of the source and the target language and the attitude of the respective Wikipedia community, where the spectrum can go from simply translating English articles to the completely independent authoring of articles.

In the test sets for German, Russian, and Ukrainian the Google translations of the 1000 English words had been manually corrected by native speakers of the respective language. As this manual work is some hindrance when exploring new languages, the question occurs whether an evaluation using the uncorrected Google translations, would also be of some use. Roughly speaking, according to the native speakers for these languages in the order of 10% of the Google translations had been erroneous, so we might also expect a drop of accuracy in this order. Table 8 shows the respective results. The expected drop is noticeable in all three cases, although its degree varies. Nevertheless the uncorrected Google translations seem suitable to give at least a rough idea of performance.

	DE→ EN	RU→ EN	UK→ EN
KW	948	861	777
1	316	319	220
2	38	44	24
3	13	15	12
4	5	10	7
5	9	5	2
6	2	3	2
7	3	1	4
8	1	1	1
9	1	1	1
10	0	1	1

Table 8: Results for three language pairs where uncorrected Google translations are used as gold standard.

Based on this observation, for the remaining languages to be considered in this paper we conducted an evaluation using a gold standard of uncorrected Google translations.

Table 9 shows the results. As can be seen in conjunction with Table 8, the Romanic languages obtain considerably better results than the Germanic or Slavic ones, and – not too surprisingly due to its high degree of word ambiguity – Chinese is the most difficult language to deal with.[7]

	ES→ EN	FR→ EN	NL→ EN	PT→ EN	ZH→ EN
KW	805	962	829	880	942
1	473	428	348	428	130
2	45	43	39	36	13
3	14	17	17	10	4
4	5	10	4	5	6
5	2	6	5	4	4
6	0	4	6	1	0
7	1	4	2	1	2
8	1	2	1	1	1
9	0	0	5	2	1
10	1	0	1	1	0

Table 9: Results for further language pairs where uncorrected Google translations are used as gold standard.

5. Conclusions and future work

We have presented a method for identifying word translations using comparable documents. Although it does not require a seed lexicon it delivers competitive results. As has been shown its knowledge poor approach can be easily applied to other language pairs, with reasonable results. Other than word segmentation and lemmatization no adaptation was required for the new language pairs, and no optimization was conducted. The quantitative evaluations are based on a gold standard which had been developed independently before the simulations were conducted.

A disadvantage of our method is that it presupposes that the alignments of the comparable documents are known. On the other hand, there are methods for finding such alignments automatically not only in special cases such as Wikipedia and newspaper texts, but also in the case of unstructured texts (although these methods may require a seed lexicon).

Our future work will concentrate on this, but also on refining the method and extending it to multiword units and further languages.

[7] For better looking results, for Chinese an evaluation method taking into account multiple translation possibilities might be desirable. On the other hand, (similar to BLEU scores in machine translation) it is better not to take these accuracy figures as absolute, but instead as a means for comparing the performances of different algorithms. We think that for this application it is preferable to consider only the most salient translations, because this way the degree of arbitrariness (as inherent in the production of any gold standard) is minimized.

6. Acknowledgments

We would like to thank Richard Forsyth for interesting discussions and for contributing most of the text of section 3.1, and for conducting much of the work described therein. The research leading to these results has received funding from the European Community's Seventh Framework Programme via the projects HyghTra (grant agreement no. 251534) and TTC (grant agreement no. 248005).

7. References

Brants T. (2000). TnT − a statistical part-of-speech tagger. *Proceedings of the 6th Applied Natural Language Processing Conference*, 224–231

Jongejan B.; Dalianis H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 145 –153.

Laws, F.; Michelbacher, L.; Dorow, B.; Scheible, C.; Heid, U.; Schütze, H. (2010). A linguistically grounded graph model for bilingual lexicon extraction. *Proceedings of Coling, Poster Volume*, 614–622.

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

Rapp, R. (1996). *Die Berechnung von Assoziatonen*. Hildesheim: Olms.

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

Rayson, P.; Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing corpora (WCC '00) - Volume 9*, 1–6.

Rumelhart, D.E.; McClelland, J.L. (1987). *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 1: Foundations*. MIT Press.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *International Conference on New Methods in Language Processing*, 44–49.

Sharoff, S.; Kopotev, M.; Erjavec, T.; Feldman, A.; Divjak, D. (2008). Designing and evaluating a Russian tagset. *Proceedings of the Sixth Language Resources and Evaluation Conference, LREC 2008*, Marrakech, 279–285.