A Methodology for Bilingual Lexicon Extraction from Comparable Corpora

Reinhard Rapp
University of Mainz, FTSK
An der Hochschule 2
D-76726 Germersheim
reinhardrapp@gmx.de

Abstract

Dictionary extraction using parallel corpora is well established. However, for many language pairs parallel corpora are a scarce resource which is why in the current work we discuss methods for dictionary extraction from comparable corpora. Hereby the aim is to push the boundaries of current approaches, which typically utilize correlations between co-occurrence patterns across languages, in several ways: 1) Eliminating the need for initial lexicons by using a bootstrapping approach which only requires a few seed translations. 2) Implementing a new approach which first establishes alignments between comparable documents across languages, and then computes cross-lingual alignments between words and multiword-units. 3) Improving the quality of computed word translations by applying an interlingua approach, which, by relying on several pivot languages, allows an effective multi-dimensional cross-check. 4) We investigate that, by looking at foreign citations, language translations can even be derived from a single monolingual text corpus.

1 Introduction

The aim of this paper is to suggest new methods for automatically extracting bilingual dictionaries, i.e. dictionaries listing all possible translations of words and multiword units, from comparable corpora. With comparable corpora we mean sets of text collections which cover roughly the same subject area in different languages or dialects, but which are not translations of each other. Their main advantages are that they don't have a translation bias (as there is no source language which could show through) and are available in by far larger quantities and for more domains than parallel corpora, i.e. collections of translated texts.

The systems to be developed are supposed to have virtually no prior knowledge on word translations. Instead, they induce this knowledge statistically using an extension of Harris' distributional hypothesis (Harris, 1954) to the multilingual case. The distributional hypothesis states that words occurring in similar contexts have related meanings. Its application led to excellent automatically created monolingual thesauri of related words. Our extension of Harris' distributional hypothesis to the multilingual case claims that the translations of words with related meanings will also have related meanings. From this it can be inferred that if two words co-occur more frequently than expected in a corpus of one language, then their translations into another language will also co-occur more frequently than expected in a comparable corpus of this other language. This is the primary statistical clue which is the basis for our work. Starting from this our aim is to develop a methodology which is capable of deriving good quality bilingual dictionaries in a language independent fashion, i.e. which can be applied to all language pairs where comparable corpora are available. In future work, to exemplify the results achieved with this method, we will generate large dictionaries comprising single words and multiword units for the following language pairs: English–German; English–French; English–Spanish; German–French; German–Spanish; French–Spanish, German–Dutch, and Spanish–Dutch.

Bilingual dictionaries are an indispensable resource for both human and machine translation. For

Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra), pages 46–55,
Beijing, China, July 31, 2015.

this reason, in the field of lexicography a lot of effort has been put into producing high quality dictionaries. For example, in rule-based machine translation producing the dictionaries is usually by far the most time consuming and expensive part of system development. But as the dictionaries are crucial in ensuring high coverage and high translation quality, a lot of effort has to be invested into them, and there are many examples where the manual creation of comprehensive dictionaries has been an ongoing process over several decades. Now that some high quality dictionaries exist, why do we suggest further research in this field? The reasons are manifold:

1) High quality dictionaries are only available for a few hundred common language pairs, usually involving some of the major European and Asian languages. But there exist about 7000 languages worldwide (Gordon & Grimes, 2005; Katzner, 2002), of which 600 have a written form. In the interest of speakers or learners of lesser used languages, at least for all possible pairs of written languages high quality dictionaries would be desirable, which means a total of 600 * (600 – 1) = 359,400 translation directions. But in practice this is impossible for reasons of time, effort, and cost. So the companies working in the field tend to concentrate on their major markets only.
2) The usage and meanings of words are adapted and modified in language of specialized domains and genres. To give an example, the word *memory* is used differently in the life sciences and in computer science. This means that in principle for each domain specific dictionaries would be desirable. Again, for a few common language pairs and commercially important subject areas such as medicine or engineering such dictionaries have been developed. But if we (conservatively) assumed only 20 subject areas, the total number of required dictionaries increases from 359,400 to 143,988,000.
3) Languages evolve over time. New topics and disciplines require the creation or borrowing (e.g. from English) of new terms (a good example is mobile computing), other terms become obsolete. This means that we cannot create our dictionaries once and forever, but need to constantly track these changes, for all language pairs, and for all subject areas.
4) Even if some companies such as specialized publishing houses (e.g. *Collins* and *Oxford University Press*), translation companies (e.g. *Systran* and *SDL*) or global players (e.g. *Google* and *Microsoft*) can afford to compile dictionaries for some markets, these dictionaries are proprietary and often not available for other companies, institutions, academia, and individuals. This is an obstacle for the advancement of the field.

Given this situation, it would be desirable to be able to generate dictionaries ad hoc as we need them from corpora of the text types we are interested in. So a lot of thought has been spent on how to produce bilingual dictionaries more efficiently than manually in the traditional lexicographic way. From these efforts, two major straits of research arose: The first is based on the exploitation of parallel corpora, i.e. collections of translated documents, as suggested by Brown et al. (1990 and 1993) in their seminal papers. They automatically extracted a bilingual dictionary from a large parallel corpus of English–French Canadian parliamentary proceedings, and then built a machine translation system around this. The development of such systems has not been without setbacks, but finally, after 15 years of research, it led to a revolution in machine translation technology and provided the basis for machine translation systems such as *Moses*, *Google Translate* and Microsoft's *Bing Translator* which are used by millions of people worldwide every day.

The second strait of research is based on comparable rather than parallel corpora. It was first suggested by Fung (1995) and Rapp (1995). The motivation was that parallel corpora are a scarce resource for most language pairs and subject areas, and that human performance in second language acquisition and in translation shows that there must be a way of crossing the language barrier that does not require the reception of large amounts of translated texts. We suggest here to replace parallel by comparable corpora. Comparable (written or spoken) corpora are far more abundant than parallel corpora, thus offering the chance to overcome the data acquisition bottleneck. This is particularly true as, given n languages to be considered, n comparable corpora will suffice. In contrast, with parallel corpora, unless translations of the same text are available in several languages, the number of required corpora c increases quadratically with the number of languages as $c = (n^2 – n)/2$.

However, the problem with comparable corpora is that it is much harder to extract a bilingual dictionary from comparable corpora than from parallel corpora. As a consequence, despite intensive research carried out over two decades (to a good part taking place in international projects such as *AC-*

CURAT, *HyghTra*, *PRESEMT*, METIS, *Kelly*, and *TTC*) no commercial breakthrough has yet been possible.

However, we feel that in recent years some remarkable improvements were suggested (e.g. dictionary extraction from aligned comparable documents and dictionary verification using cross checks based on pivot languages). They cannot solve the problem when used in isolation, but when amended and combined they may well have the potential to lead to substantial improvements. In this paper we try to come up with a roadmap for this.

2 Methodology

Although, if at all, it is more likely that the mechanisms underlying human second language acquisition are based on the processing of comparable rather than parallel corpora, we do not attempt to simulate the complexities of human second language acquisition. Instead we argue that it is possible by purely technical means to automatically extract information on word- and multiword-translations from comparable corpora. The aim is to push the boundaries of current approaches, which often utilize similarities between co-occurrence patterns across languages, in several ways:

1. Eliminating the need for initial dictionaries.
2. Looking at aligned comparable documents rather than at comparable corpora.
3. Utilizing multiple pivot languages in order to improve dictionary quality.
4. Considering word senses rather than words in order to solve the ambiguity problem.
5. Investigate in how far foreign citations in monolingual corpora are useful for dictionary generation.
6. Generating dictionaries of multiword units.
7. Applying the approach to different text types.
8. Developing a standard test set for evaluation.

Let us now look point by point at the above list of research objectives with an emphasis on methodological and innovative aspects.

2.1 Eliminating the need for initial dictionaries

The standard approach for the generation of dictionaries using comparable corpora operates in three steps: 1) In the source language, find the words frequently co-occurring with a given word whose translation is to be determined. 2) Translate these frequently co-occurring words into the target language using an initial dictionary. 3) In the target language, find the word which most frequently co-occurs with these translations.

There are two major problems with this approach: Firstly, an already relatively comprehensive initial dictionary of typically more than 10,000 entries (Rapp, 1999) is required which will often be a problem for language pairs involving lesser used languages or when existing dictionaries are copyright protected or not available in machine readable form. Secondly, depending on the coverage of this dictionary, quite a few of the requested translations may not be known. For these reasons a method not requiring an initial dictionary would be desirable. Let us therefore outline our proposal for a novel bootstrapping approach which requires only a few seed translations. The underlying idea is based on multi-stimulus associations (Rapp, 1996; Rapp, 2008; Lafourcade & Zampa, 2009; Rapp & Zock, 2014). There is also related work in cognitive science. It often goes under the label of the *remote association test*, but essentially pursues the same ideas (Smith et al., 2013).

As experience tells, associations to several stimuli are non-random. For example, if we present the word pair *circus – laugh* to test persons and ask for their spontaneous associations, a typical answer will be *clown*. Likewise, if we present *King – daughter*, many will respond with *princess*. Like the associative responses to single words, the associative answers to pairs of stimuli can also be predicted with high precision by looking at the co-occurrences of words in text corpora. A nice feature about the word pair associations is that the number of possible word pairs increases with the square of the vocabulary size considered. For a vocabulary of n words, the number of possible pairwise combinations (and likewise the number of associations) is $n * (n - 1) / 2$. This means that for a vocabulary of 10 words we have 45, for a vocabulary of 100 words we have 4,950, and for a vocabulary of 1000 words we have 499,500 possible word pairs, and each of these pairs provides valuable information.[1]

[1] As will become later on, this is actually one of the rare cases where large numbers work in favour of us, thus making the method well suited for the suggested bootstrapping approach. This behavior is in contrast to most other applications in natural language processing. For example, in syntax parsing or in machine translation the number of possible parse trees or sentence translations tends to grow exponentially with the length of a sentence. But the higher the number of possibilities, the more difficult it gets to filter out the correct variant.

To exemplify the suggested approach, let us assume that our starting vocabulary consists of the four words *circus*, *laugh*, *King*, and *daughter*. We assume that their translations into the target language are known. If our target language is German, the translations are *Zirkus*, *lachen*, *König*, and *Tochter*. Separately for the source and the target language, based on corpus evidence we compute the multi-stimulus associations for all possible word pairs (compare Rapp, 2008):

English:
circus – laugh → clown
circus – King → lion
circus – daughter → artiste
laugh – King → jester
laugh – daughter → joke
King – daughter → princess

German:
Zirkus – lachen → Clown
Zirkus – König → Löwe
Zirkus – Tochter → Artistin
Lachen – König → Hofnarr
lachen – Tochter → Witz
König – Tochter → Prinzessin

Now our basic assumption is that the corresponding English and German multi-stimulus associations are translations of each other. This means that to our initial four seed translations we can now add a further six newly acquired translations, namely *clown → Clown*, *lion → Löwe*, *artiste → Artistin*, *jester → Hofnarr*, *joke → Witz*, *princess → Prinzessin*. Together with the four seed translations, this gives us a total of ten known translations. With these ten translations we can restart the process, this time with a much higher number of possible pairs (45 pairs of which 35 are new). Once this step is completed, ideally we would have 45 * (45 – 1) / 2 = 990 known translations. In continuation, with a few more iterations we cover a very large vocabulary.

Of course, for the purpose of demonstrating the approach we have idealized matters here. In reality, many word pairs will not have salient associations, so the associations which we compute can be somewhat arbitrary. This means that our underlying assumption, namely that word pair associations are equivalent across languages, may not hold for non-salient cases, and even when the associations are salient there can still be discrepancies caused by cultural, domain-dependent and other differences. For example, the word pair *pork – eat* might evoke the association *lunch* in one culture, but *forbidden* in another. But non-salient associations can be identified and eliminated by applying a significance test on the measured association strengths. And cultural differences are likely to be small in comparison to the commonalities of human life as expressed through language. Would this not be true, it should be almost impossible to translate between languages with different cultural backgrounds, but experience tells us that this is still possible (though more difficult).

It should also be noted that the suggested approach, like most statistical approaches used in NLP, should show a great deal of error tolerance. The iterative process should converge as long as the majority of computed translations is correct. Also, the associative methodology implies that incorrect translations will typically be caused by mixups between closely related words, which will limit the overall negative effect.

If required to ensure convergence, we can add further levels of sophistication such as the following: a) Compute salient associations not only for word pairs, but also for word triplets (e.g. *pork – eat – Muslim → forbidden*; *pork – eat – Christian → ok*). b) Use translation probabilities rather than binary yes/no decisions. c) Use pivot languages to verify the correctness of the computed translations (see section 2.4 below). d) Look at aligned comparable documents (see below).

2.2 Looking at aligned comparable documents rather than at comparable corpora

Here we investigate an alternative approach to the above. It also does not require a seed lexicon, but instead has higher demands concerning the comparable corpora to be used. For this approach the comparable corpora need to be alignable at the document level, i.e. it must be possible to identify correspondences between the documents in two comparable corpora of different languages. This is straightforward e.g. for Wikipedia articles where the so-called interlanguage links (created manually by the authors) connect articles across languages. But there are many more common text types which are easily alignable, among them newspaper corpora where the date of publication gives an important clue, or scientific papers whose topics tend to be so narrow that a few specific internationalisms or proper names can be sufficient to identify the correspondences.

Once the alignment at the document level has been conducted, the next step is to identify the most salient keywords in each of the documents. There are a number of well established ways of doing so, among them Paul Rayson's method of comparing

the observed term frequencies in a document to the average frequencies in a reference corpus using the log-likelihood ratio, or – alternatively – the Likely-system as developed by Paukkeri & Honkela (2010). By applying these keyword extraction methods the aligned comparable documents are converted to aligned lists of keywords. Some important properties of these lists of aligned keywords are similar to those of aligned parallel sentences, which means that there is a chance to successfully apply the established statistical machinery developed for parallel sentences. We conducted a pilot study using a self-developed robust alternative to GIZA++, with promising results (Rapp, Sharoff & Babych, 2012). In principle, the method is applicable not only to the problem of identifying the translations of single words, but also of identifying the translations of multiword units, see section 2.6 below.

2.3 Utilizing multiple pivot languages in order to improve dictionary quality

We propose to systematically explore the possibility of utilizing the dictionaries' property of *transitivity*. What we mean by this is the following: If we have two dictionaries, one translating from language A to language B, the other from language B to language C, then we can also translate from A to C by using B as the pivot language (also referred to as bridge language, intermediate language, or interlingua). That is, the property of transitivity, although having some limitations due to the ambiguity problem, can be exploited for the automatic generation of a raw dictionary with mappings from A to C. On first glance, one might consider this unnecessary as our corpus-based approach allows us to generate such a dictionary with higher accuracy directly from the respective comparable corpora.

However, the above implies that we have now two ways of generating a dictionary for a particular language pair, which means that in principle we can validate one with the other. Furthermore, given several languages, there is not only one method to generate a transitivity-based dictionary for A to C, but there are several. This means that by increasing the number of languages we also increase the possibilities of mutual cross-validation. In this way a highly effective multi-dimensional cross-check can be realized.

Utilizing transitivity is a well established technique in manual dictionary lookup when people interested in uncommon language pairs (where no dictionary is available) use two dictionaries involving a common pivot language. Likewise, lexicographers often use this concept when manually creating dictionaries for new language pairs based on existing ones. However, this has not yet been explored at a large scale in a setting like ours. We propose to use many pivot languages in parallel, and to introduce a voting system where a potential translation of a source word is ranked according to the number of successful cross-validations.

2.4 Considering word senses rather than words in order to solve the ambiguity problem

As in natural language most words are ambiguous, and as the translation of a word tends to be ambiguous in a different way than the original source language word (especially if we look at unrelated languages belonging to different language families), our extension of Harris's distributional hypothesis which says that the translations of two related words should be related again (see Section 1) is only an approximation but not strictly applicable. But in principle it would be strictly applicable and therefore lead to better results if we conducted a word sense disambiguation on our comparable corpora beforehand. Hereby we assume that the sense inventories for the languages to be considered are similar in granularity and content.[2] We therefore propose to sense disambiguate the corpora, and to apply our method for identifying word translations on the senses. As a result, we will not only obtain a bilingual dictionary, but also an alignment of the two sense inventories.

As versions of *WordNet* are available for all five languages mentioned in Section 1 (English, French, German, Spanish, Dutch), we intend to use these WordNets as our sense inventories. Regarding some criticism that they are often too fine grained for practical applications (Navigli, 2009), we will consider attempts to automatically derive more coarse-grained sense inventories from them (Navigli et al., 2007). Given the resulting sense inventories, we will apply an open source word sense disambiguation algorithm such as Ted Pedersen's *SenseRelate* software (alternatives are *BabelNet*, *UKB* and other systems as e.g. used in the SemEval word sense disambiguation competitions).

Relying on the WordNet senses means that the methodology is not applicable to languages where a version of WordNet is not available. As this is a serious shortcoming, we have looked at methods for generating corpus-specific sense inventories in an unsupervised way (Pantel & Lin, 2002; Bordag,

[2] Similar sense inventories across languages can be expected under the assumption that the senses reflect observations in the real world.

2006; Rapp, 2003; SemEval 2007 and 2010 task "Word sense induction"). In an attempt to come up with an improved algorithm, we propose a novel bootstrapping approach which conducts word sense induction and word sense disambiguation in an integrated fashion. It starts by tagging each content word in a corpus with the strongest association that occurs nearby. For example, in the sentence "*He gets money from the bank*", the word *bank* would be tagged with *money* as this is the strongest association occurring in this neighborhood. Let us use the notation [bank < money] to indicate this. From the tagged corpus a standard distributional thesaurus is derived (Pantel & Lin, 2002). This thesaurus would, for example, show that [bank < money] is closely related to [bank < account], but not to [bank < river]. For this reason, all occurrences of [bank < money] and [bank < account] would be replaced by [bank < money, account], but [bank < river] would remain unchanged. Likewise for all other strongly related word/tag combinations. Subsequently, in a second iteration a new distributional thesaurus is computed, leading to further mergers of word/tag combinations. This iterative process is to be repeated until there are no more strong similarities between any entries of a newly created thesaurus. At this point the result is a fully sense tagged corpus where the granularity of the senses can be controlled as it depends on the similarity threshold used for merging thesaurus entries.

2.5 Investigating in how far foreign citations in monolingual corpora can be utilized for dictionary generation

Traditional foreign language teaching, where the teacher explains the foreign language using the native tongue of the students, has often been criticized. But there can be no doubt that it works at least to some extent. Apparently, the language mix used in such a teaching environment is non-random, which is why we start from the hypothesis that it should be possible to draw conclusions on word translations given a corpus of such classroom transcripts. We suggest that the translations of words can be discovered by looking at strong associations between the words of the teaching language and the words of the foreign language. In a 2nd-language teaching environment the words of the foreign language tend to be explained using corresponding words from the teaching language, i.e. these two types of words tend to co-occur more often than to be expected by chance.

However, as it is not easy to compile transcripts of such classroom communications in large enough quantities, we assume that the use of foreign language citations in large newspaper or web corpora follows similar principles (for a pilot study see Rapp & Zock, 2010b). The following two citations from the Brown Corpus (Francis & Kuçera, 1989) are meant to provide some evidence for this (underscores by us):

1. The tables include those for the classification angles , refractive indices , and melting points of the various types of crystals . Part 2 of Volume /1 , and Parts 2 and 3 of Volume /2 , contain the crystal descriptions . These are grouped into sections according to the crystal system , and within each section compounds are arranged in the same order as in Groth 's CHEMISCHE KRYSTALLOGRAPHIE . An alphabetical list of chemical and mineralogical names with reference numbers enables one to find a particular crystal description . References to the data sources are given in the crystal descriptions .

2. On the right window , at eye level , in smaller print but also in gold , was Gonzalez , Prop. , and under that , Se Habla Espanol . Mr. Phillips took a razor to Gonzalez , Prop. , but left the promise that Spanish would be understood because he thought it meant that Spanish clientele would be welcome .

In the first example, the German book title "Chemische Krystallographie"[3] (meaning *Chemical Crystallography*) is cited. In its context the word *chemical* occurs once and the word forms *crystal* and *crystals* occur five times. In the second example, the phrase "Se Habla Espanol" is cited (meaning: *Spanish spoken* or *We speak Spanish*), and in its context we find "Spanish would be understood" which comes close to a translation of this phrase. (And a few words further in the same sentence the word "Spanish" occurs again.)

Although foreign language citations are usually scarce in standard corpora, lexicon extraction from monolingual corpora should still be feasible for heavily cited languages such as English. For other languages lexicon construction should be possible via pivot languages, see Section 2.3 above. The problem that the same word form can occur in several languages but with different meanings (called "homograph trap" in Rapp & Zock, 2010b) can be approached by looking at several source languages at the same time and by eliminating interpretations which are not consistent with several of the languages. We intend to apply this method to all language pairs, and use it in a supplementary fashion to enhance the other approaches. This looks prom-

[3] Note that *Krystallographie* is an old spelling. The modern spelling is *Kristallographie*.

ising as the method provides independent statistical clues from a different type of source.

2.6 Generating dictionaries of multiword units

Due to their sheer numbers, the treatment of multiword units is a weakness of traditional lexicography. Whereas a reasonable coverage of single words may require in the order of 100,000 dictionary entries, the creation of multiword units is highly productive so that their number can be orders of magnitude higher, making it infeasible to achieve good coverage using manual methods. In contrast, most automatic methods for dictionary extraction, including the ones described above, can be applied to multiword units in a straightforward way. The only prerequisite is that the multiword units need to be known beforehand, that is, in a pre-processing step they must be identified and tagged as such in the corpora. There exist numerous methods for this, most of them relying on measures of mutual information between neighbouring words (e.g. Smadja, 1993; Paukkeri & Honkela, 2010). Our intention is to adopt the language independent "Likely" system for this purpose (Paukkeri & Honkela, 2010). Using the methods described in Sections 2.1 to 2.5, we will generate dictionaries of multiword units for all language pairs considered, i.e. involving English, French, German, Spanish, and Dutch, and then evaluate the dictionaries as outlined in section 2.8.

Our expectation is that the problem of word ambiguity will be less severe with multiword units than it is with single words. There are two reasons for this, which are probably two sides of the same medal: One is that rare words tend to be less ambiguous than frequent words, as apparently in human language acquisition a minimum number of observations is required to learn a reading, and the chances to reach this minimum number are lower for rare words. As multiword units are less frequent than their rarest constituents, on average their frequencies are lower than the frequencies of single words. Therefore it can be expected that they must be less ambiguous on average. The other explanation is that in multiword units the constituents tend to disambiguate each other, so fewer readings remain.

2.7 Applying the approach to different text types

By their nature, the dictionaries generated using the above algorithms will always reflect the contents of the underlying corpora, i.e. their genre and topic. This means that if the corpora consist of newspaper articles on politics, the generated dictionaries will reflect this use of language, and likewise with other genres and topics. It is of interest to investigate these effects. However, as for a reasonable coverage and quality of the extracted dictionaries we need large corpora (e.g. larger than 50 million words) for all five languages, we feel that for a first study it is only realistic to make just a few rough distinctions in a somewhat opportunistic way: a) newspaper articles; b) parliamentary proceedings; c) encyclopaedic articles; d) general web documents. The resulting dictionaries will be compared qualitatively and quantitatively. However, in the longer term it will of course be of interest to aim for more fine-grained distinctions of genre and topic.

2.8 Developing a standard test set for evaluation

As previous evaluations of the dictionary extraction task were usually conducted with ad hoc test sets and thus were not comparable, Laws et al. (2010) noted an urgent need for standard test sets. In response to this, we intend to work out and publish a gold standard which covers all of our eight language pairs and will ensure that words of a wide range of frequencies are appropriately represented. All results on single words are to be evaluated using this test set.

Little work has been done so far on multiword dictionary extraction using comparable corpora (an exception is Rapp & Sharoff, 2010), and no widely accepted gold standard exists. A problem is that there are many ways how to define multiword units. To explore these and to provide for different needs, we aim for five types of test sets of at least 5000 multiword units and their translations. The test sets are to be generated semi-automatically in the following ways:

a) Multiword units connected by Wikipedia interlanguage links.
b) Multiword units extracted from a parallel corpus which was word-aligned using GIZA++.
c) Multiword units extracted from phrase tables as generated using the Moses toolkit.
d) Multiword units extracted with a co-occurrence based system such as *Likely* (Paukkeri & Honkela, 2010) and redundantly translated with several translation systems, using voting to select translations.
e) Multiword named entities taken from *JRC-Names* (as provided by the European Commission's Joint Research Centre).

The results on the multiword dictionary extraction task are to be evaluated using each of these gold standards.

3 Discussion

In this section we discuss the relationship of the suggested work to the state of the art of research in the field. Hereby we concentrate on how the previous literature relates to the eight subtopics listed above. A more comprehensive survey of the field of bilingual dictionary extraction from comparable corpora can be found in Sharoff et al. (2013).

1) *Eliminating the need for initial dictionaries:* This problem has been approached e.g. by Rapp (1995), Diab & Finch (2000), Haghighi et al. (2008), and Vulic & Moens (2012). None of the suggested solutions seems to work well enough for most practical purposes. Through its multilevel approach, the above methodology aims to achieve this.
2) *Looking at aligned comparable documents rather than at comparable corpora:* Previous publications concerning this are Schafer & Yarowsky (2002), Hassan & Mihalcea (2009), Prochasson & Fung (2011) and Rapp et al. (2012). In our view, the full potential has not yet been unveiled.
3) *Utilizing multiple pivot languages in order to improve dictionary quality:* The (to our knowledge) only previous study in such a context was conducted by ourselves (Rapp & Zock, 2010a), and uses only a single pivot language. In contrast, here we suggest to take advantage of multiple pivot languages.[4]
4) *Considering word senses rather than words in order to solve the ambiguity problem:* Gaussier et al. (2004) use a geometric view to decompose the word vectors according to their senses. In contrast, we will use explicit word sense disambiguation based on the WordNet sense inventory. Annotations consistent with human intuitions are easier to verify and thus the system can be better optimized.
5) *Investigating in how far foreign citations in monolingual corpora can be used for dictionary generation:* To our knowledge, apart from our own (see Rapp & Zock, 2010b) there is no other previous work on this.
6) *Generating dictionaries of multiword units:* Robitaille et al. (2006) and the TTC project (http://www.ttc-project.eu/) dealt with this in a comparable corpora setting but did not make their results available. In contrast, the intention here is to publish the full dictionaries.
7) *Applying the approach to different text types:* Although different researchers used a multitude of comparable corpora, to our knowledge there exists no systematic comparative study concerning different text types in the field of bilingual dictionary extraction.
8) *Developing a standard test set for evaluation:* Laws et al. (2010) pointed out the need for a common test set and provided one for the language pair English – German. Otherwise in most cases ad hoc test sets were used, and to our knowledge no readily available test set exists for multiword units.

4 Conclusions

A core problem in NLP is the problem of ambiguity in a multilingual setting. Entities in natural language tend to be ambiguous but can be interpreted as mixtures of some underlying unambiguous entities (e.g. a word's senses). The problem in simulating, understanding, and translating natural language is that we can only observe and study the complicated behavior of the ambiguous entities, whereas the presumably simpler behavior of the underlying unambiguous entities remains hidden. The proposed work shows a way how to deal with this problem. This is relevant to most other fields in natural language processing where the ambiguity problem is also of central importance, such as MT, question answering, text summarization, thesaurus construction, information retrieval, information extraction, text classification, text data mining, speech recognition, and the semantic web.

The suggested work investigates new methods for the automatic construction of bilingual dictionaries, which are a fundamental resource in human translation, second language acquisition, and machine translation. If approached in the traditional lexicographic way, the creation of such resources has often taken years of manual work involving numerous subjective and potentially controversial decisions. In the suggested framework, these human intuitions are replaced by automatic processes which are based on corpus evidence. The developed systems will be largely language independent and will be applied to eight project language pairs involving the five European languages mentioned in Section 1. The suggested approach is of interest

[4] We use here the term *pivot language* as the potentially alternative term *bridge language* is used by Schafer &Yarowsky (2002) in a different sense, relating to orthographic similarities.

as the recent advances in the field concerning e.g. bootstrapping algorithms, alignment of comparable documents, and word sense disambiguation were conducted in isolated studies but need to be amended, combined, and integrated into a single system.

For human translation, by preparing the resulting dictionaries in XML they can be made compatible for use with standard Translation Memory systems. This way they are available for professional translation (especially in technical translation, technical writing, and interpreting) where there is a high demand in specialized dictionaries, thus supporting globalization and internationalization.

The work is also of interest from a cognitive perspective, as a bilingual dictionary can be seen as a collection of human intuitions across languages. The question is if these intuitions do find their counterpart in corpus evidence. Should this be the case, this would support the view that human language acquisition can be explained by unsupervised learning on the basis of perceived spoken and written language. If not, other sources of information available for language learning would have to be identified, which may, for example, include an equivalent of Chomsky's language acquisition device.

Acknowledgments

This research was supported by a Marie Curie Career Integration Grant within the 7th European Community Framework Programme. I would like to thank Silvia Hansen-Schirra for her support of this work and valuable comments.

References

Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S. (1990). A statistical approach to machine translation. *Computational Linguistics* 16(2), 79–85.

Brown, P., Pietra, S.D., Pietra, V.D., Mercer, R. (1993): The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* 19(2), 263–312.

Bordag, S. (2006). Word sense induction: triplet-based clustering and automatic evaluation. *Proceedings of EACL 2006*, Trento, Italy. 137–144.

Diab, M., Finch, S. (2000): A statistical wordlevel translation model for comparable corpora. *Proceedings of the Conference on Content-Based Multimedia Information Access* (RIAO).

Francis, W. Nelson; Kuçera, Henry (1989). *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers.* Providence, R.I.: Brown University, Department of Linguistics.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. *Proceedings of the Third Annual Workshop on Very Large Corpora*, Boston, Massachusetts. 173–183.

Gaussier, E., Renders, J.M., Matveeva, I., Goutte, C., Djean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. *Proceedings of the 42nd ACL*, Barcelona, Spain, 526–533.

Gordon, R. G.; Grimes, B. F. (eds.) (2005). *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, 15th edition.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., Klein, D. (2008): Learning bilingual lexicons from monolingual corpora. *Proceedings of ACL-HLT 2008*, Columbus, Ohio. 771–779.

Harris, Z.S. (1954). Distributional structure. *Word*, 10(23), 146–162.

Hassan, S., Mihalcea, R. (2009): Cross-lingual semantic relatedness using encyclopedic knowledge. *Proceedings of EMNLP*.

Katzner, K. (2002). *The Languages of the World*. Routledge, London/New York, 3rd edition.

Lafourcade, M.; Zampa, V. (2009). JeuxDeMots and PtiClic: games for vocabulary assessment and lexical acquisition. *Proceedings of Computer Games, Multimedia & Allied Technology 09* (CGAT'09). Singapore.

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 2010*, Beijing, China.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, Vol. 41, No. 2, Article 10.

Navigli, R.; Litkowski, K.; Hargraves, O. (2007). SemEval-2007 task 07: Coarse-grained English all-words task. *Proceedings of the Semeval-2007 Workshop at ACL 2007*, Prague, Czech Republic.

Pantel, P.; Lin, D. (2002). Discovering word senses from text. *Proceedings of ACM SIGKDD*, Edmonton, 613–619.

Paukkeri, M.-S.; Honkela, T. (2010). Likey: unsupervised language-independent keyphrase extraction. *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval) at ACL 2012*, 162–165.

Prochasson, E., Fung, P. (2011). Rare word translation extraction from aligned comparable documents. *Proceedings of ACL-HLT*, Portland.

Rapp, R. (1995). Identifying word translations in non-parallel texts. *Proceedings of the 33rd ACL*, Cambridge, MA, 320–322.

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

Rapp, R. (1999): Automatic identification of word translations from unrelated English and German corpora. *Proceedings of the 37th ACL*, Maryland, 395–398.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. *Proceedings of the Ninth Machine Translation Summit*, 315–322.

Rapp, R. (2008). The computation of associative responses to multiword stimuli. *Proceedings of the Workshop on Cognitive Aspects of the Lexicon* (COGALEX) at Coling 2008, Manchester. 102–109.

Rapp, R.; Sharoff, S. (2014). Extracting multiword translations from aligned comparable documents. *Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra).* Gothenburg, Sweden.

Rapp, R., Sharoff, S., Babych, B. (2012). Identifying word translations from comparable documents without a seed lexicon. *Proceedings of the 8th Language Resources and Evaluation Conference, LREC 2012*, Istanbul, Turkey.

Rapp, R., Zock, M. (2010a). Automatic dictionary expansion using non-parallel corpora. In: Fink, A., Lausen, B., Ultsch, W.S.A. (eds.): *Advances in Data Analysis, Data Handling and Business Intelligence. Proceedings of the 32nd Annual Meeting of the GfKl*, 2008. Springer, Heidelberg.

Rapp, R., Zock, M. (2010b): The noisier the better: Identifying multilingual word translations using a single monolingual corpus. *Proceedings of the 4th International Workshop on Cross Lingual Information Access*, COLING 2010, Beijing, 16–25.

Rapp, R.; Zock, M. (2014). The CogALex-IV Shared Task on the Lexical Access Problem. *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, COLING 2014, Dublin, Ireland, 1–14.

Robitaille, X., Sasaki, Y., Tonoike, M., Sato, S., Utsuro, T. (2006). Compiling French-Japanese terminologies from the web. *Proceedings of the 11th Conference of EACL*, Trento, Italy, 225–232.

Schafer, C., Yarowsky, D (2002).: Inducing translation lexicons via diverse similarity measures and bridge languages. *Proceedings of CoNLL.*

Sharoff, S.; Rapp, R.; Zweigenbaum, P. (2013). Overviewing important aspects of the last twenty years of research in comparable corpora. In: S. Sharoff, R. Rapp, P. Zweigenbaum, P. Fung (eds.): *Building and Using Comparable Corpora*. Heidelberg: Springer, 1–18.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics* 19 (1), 143–177.

Smith, Kevin A.; Huber, David E.; Vul, Edward (2013). Multiply-constrained semantic search in the Remote Associates Test. *Cognition* 128, 64–75.

Vulic, Ivan; Moens, Marie-Francine (2012). Detecting highly confident word translations from comparable corpora without any prior knowledge. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, 449–459.