# Semantic Relations in Bilingual Lexicons

YVES PEIRSMAN
Stanford University & QLVL, University of Leuven, Belgium
Research Foundation – Flanders (FWO)
and
SEBASTIAN PADÓ
Institute of Computational Linguistics, University of Heidelberg, Germany

Bilingual lexicons, essential to many NLP applications, can be constructed automatically on the basis of parallel or comparable corpora. In this article, we make two contributions to their induction from comparable corpora. The first one concerns the creation of these lexicons. We show that seed lexicons can be improved by adding a bootstrapping procedure that uses cross-lingual distributional similarity.

The second contribution concerns the evaluation of bilingual lexicons. It is generally based on translation lexicons, which corresponds to the implicit assumption that (cross-lingual) synonymy is the semantic relation of primary interest, even though other semantic relations like (cross-lingual) hyponymy or co-hyponymy make up a considerable portion of translation pair candidates proposed by distributional methods.

We argue that the focus on synonymy is an oversimplification and that many applications can profit from the inclusion of other semantic relations. We study what effect these semantic relations have on two cross-lingual tasks: the cross-lingual projection of polarity scores and the cross-lingual modeling of selectional preferences. We find that the presence of non-synonymous semantic relations may negatively affect the former of these tasks, but benefit the latter.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Bilingual lexicons, vector space semantics, multilingual knowledge induction, semantic relations, sentiment analysis, selectional preferences

## 1. INTRODUCTION

There is a large need in Natural Language Processing for knowledge about translational relationships between languages. Bilingual lexicons are not only crucial for cross- and multilingual tasks like machine translation or cross-lingual information retrieval, but also for cross-lingual knowledge induction, that is, the exploitation of translational knowledge to improve monolingual methods. Because the manual construction of bilingual lexicons is a labor-intensive task, a considerable body of work has focused on methods for their automatic induction. This induction can be

An earlier version of part of this work was presented at NAACL 2010. The present article adds (a) a detailed manual analysis and discussion of different semantic relations in bilingual semantic spaces and (b) introduces the cross-lingual induction of sentiment scores.

approached with a parallel corpus or comparable corpora. If a parallel corpus is available, translations are discovered on the basis of frequent alignments at the word and phrase levels, as is standard practice in Machine Translation. Since parallel corpora are scarce, however, the use of comparable corpora and the identification of translations on the basis of cross-lingual distributional similarity is an interesting proposition (see Section 2 for a discussion).

Automatically constructed bilingual lexicons are usually evaluated as models of cross-lingual synonymy, against human-constructed translational lexicons. However, it is unclear whether this is the right objective function to maximize. Monolingually, it is well-known that distributional similarity corresponds to a variety of semantic relations, and thus bilingual lexicons constructed from distributional similarity also contain other semantic relations by default. An evaluation which focuses exclusively on synonymy will treat such other relations as errors. From the perspective of cross-lingual natural language processing, this seems at best a considerable simplification, given that it is well established monolingually that computational semantics tasks can profit from knowledge about semantic relations such as hypernymy, hyponymy, or antonymy.

Our paper contributes to the development and understanding of distributional bilingual lexicons in two ways. Our first contribution is technical. We show that the initial results of cross-lingual distributional similarity can be improved by a bootstrapping procedure that benefits from knowledge about new translations that were not present in the seed lexicon. Our second, and more fundamental, contribution is a comprehensive evaluation of the semantic relations in the resulting bilingual lexicons, with an eye to a better understanding of the semantic behavior of bilingual distributional models and their use in more concrete applications. We focus on two applications in particular: the cross-lingual transfer of polarity scores and of selectional preferences. We demonstrate that selectional preference induction can draw a clear profit from non-synonymous translations.

## 2. RELATED WORK

Bilingual lexicons are traditionally induced with parallel corpora, on the basis of word alignment models [Och and Ney 2003]. However, because parallel corpora are relatively scarce resources, research is trying to approach bilingual lexicon induction on the basis of distributional information from comparable corpora. Distributional models of word meaning capture the semantic similarity and relatedness between two words as the similarity between the contexts in which they occur [Turney and Pantel 2010]. Context can be defined in a number of ways. Document-based models [Landauer and Dumais 1997] extract words that occur in similar documents, word-based models [Sahlgren 2006] look for words that often co-occur with the same context words, while syntax-based methods [Padó and Lapata 2007] identify words that frequently appear in the same syntactic relations. The definition of context influences the type of semantic relation that the models identify. Syntax-based methods are generally taken to be the best models of taxonomic similarity [Kilgarriff and Yallop 2000; Peirsman et al. 2008; Baroni and Lenci 2010], followed by word-based models that look for context words within a small context window [Sahlgren 2006]. Word-based methods with large context windows and document-based models in particular have proved to be better models of general semantic association [Sahlgren 2006; Peirsman and Geeraerts 2009].

For the construction of bilingual lexicons, these measures of distributional similarity can also be computed between words from two different languages. Rapp [1995] was one of the first to apply the distributional paradigm to two monolingual corpora. By maximizing the similarity between a German and English target-word-by-context-word matrix, he arrived at two matrices in which the order of the words in the rows and columns was the same. Due to the high computational cost of this approach, Rapp [1999] later moved on to working with a small bilingual seed lexicon to identify initial context word correspondences between English and German. In this way, individual context vectors can be compared directly. Similar lexicon-based approaches were also used by Fung and McKeown [1997] for English–Japanese, by Chiao and Zweigenbaum [2002] for French–English, and by Holmlund et al. [2005] for English–German, among other examples. In addition to context words, the induction of bilingual lexicons has benefited from orthographic clues [Haghighi et al. 2008] and multilingual dependency parses [Garera et al. 2009]. Although none of these methods yet perform as well as techniques based on parallel corpora, high rates of accuracy have been reported for a variety of language pairs, stimulating more research in this area.

The cross-lingual knowledge contained in bilingual lexicons can be put to use in several ways. Two core applications are cross-language information retrieval and cross-lingual knowledge induction. The goal of cross-language information retrieval is to identify information in a language different from that in which the query was formulated. A variety of distributional models have been used for this task, including Latent Semantic Analysis [Dumais et al. 1996] and topic models [De Smet and Moens 2009]. Query translation has been shown to contribute to information retrieval in related languages (translations from Portuguese, German, Spanish and Swedish to English in Markó et al. [2005]) as well as unrelated languages (from Japanese to English in Sadat et al. [2003]).

In cross-lingual knowledge induction, word translations between two languages are used to transfer knowledge about utterances in one language to another. This knowledge can be of any type, from part-of-speech tagging and noun phrase bracketing [Yarowsky and Ngai 2001], to parsing [Hwa et al. 2005; Zeman and Resnik 2008; Zhao et al. 2009] and lexical-semantic information like sense labels [Diab and Resnik 2002], concept distances [Mohammad et al. 2007] and verb classes [Merlo et al. 2002]. In short, bilingual lexicons can be used to support a large number of NLP applications.

## 3. CONSTRUCTING A BILINGUAL LEXICON

As discussed above, we use cross-lingual distributional similarity to construct a bilingual vector space that can be interpreted as a bilingual lexicon. This setup is a multilingual extension of the well-known distributional hypothesis [Harris 1954] and is illustrated in Figure 1(a). Let us assume that we have a semantic space whose dimensions are labeled with *bilingual context word pairs* [Rapp 1995] like *sweet/süß* or *red/rot*. We can now represent words of the two languages in the same space by gathering co-occurrence counts with these context words from two monolingual,

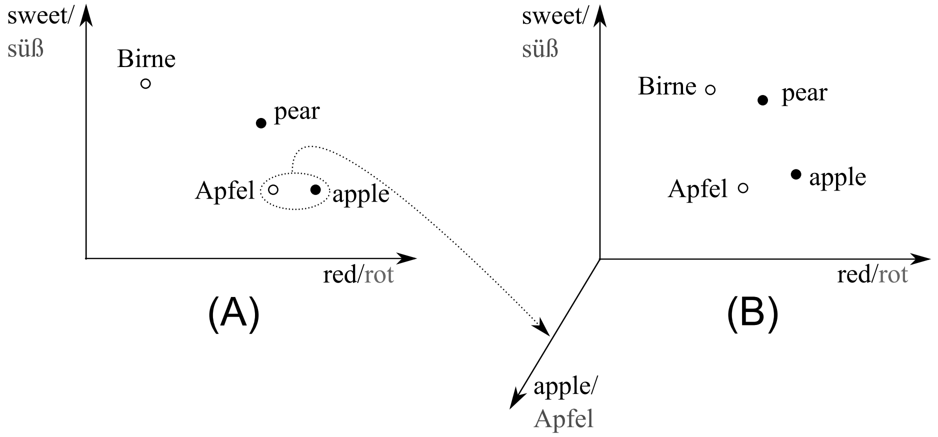

Fig. 1. Bootstrapping a bilingual vector space (example for English–German)

non-parallel corpora of either language. For example, English *apple* co-occurs with *red* more frequently than with *sweet*, and is therefore located in the bottom right quadrant of the space. In German, we see the same pattern: *Apfel* co-occurs more frequently with *rot* than with *süß*. We can now compute the cross-lingual semantic similarity of these terms, using a distributional similarity measure (see Lee [1999] for an overview). We call the set of $n$ words in the other language that are most similar to some target word the *(top) n nearest cross-lingual neighbors*, and the single nearest cross-lingual neighbor of a target word its *translation candidate*. A bilingual lexicon can be read off the space by collecting the translation candidates for all target words.

### 3.1 Bootstrapping

The bilingual space shown in Figure 1(a) has one bottleneck: to compute cross-lingual semantic similarity, it requires a set of bilingual word pairs as bilingual dimensions. Most distributional approaches indeed assume that some seed lexicon is available. We propose that a cost-efficient way to create a bilingual vector space is to complement the seed lexicon with a *bootstrapping* procedure [Fung and Yee 1998; Riloff and Shepherd 1999; Gamallo Otero 2008] that iteratively adds pairs of high-confidence translation candidates and their target words as new bilingual dimensions to the semantic space. This procedure is repeated until convergence, which in practice happens after a small number of iterations (less than 5). In this manner, the dimensionality of the space is increased, and more information about the contextual distribution of the words in the two languages is added. Bootstrapping also has the chance to correct errors which might have existed in the seed lexicon. This makes it possible to use automatically generated, noisy seed lexicons, like cognate lists [Peirsman and Padó 2010].

Figure 1, seen in its entirety, shows one step of the bootstrapping procedure. In subfigure 1(a), we find the English words *pear* and *apple*, and their German translations *Birne* and *Apfel*. In our example, *Birne* and *pear* are not very similar, which is a typical situation that can arise from any number of reasons. However, *apple* and *Apfel* are mutual translation candidates (mutual nearest cross-lingual neighbors). Thus, we can add them as a new dimension to the space, as shown in Figure 1(b). This extra dimension improves cross-lingual semantic similarity estimates, which leads to more reliable translation pairs: *Birne* and *pear* are now more similar to each other, because both co-occur with *Apfel/apple*.

A crucial step in this bootstrapping procedure is the estimation of confidence in a translation pair. Errors, particularly in the first stages, can "poison" the bootstrapping process [Riloff and Shepherd 1999]. We therefore require translation pairs to be symmetric translation pairs, that is, $w_1$ has to be the translation candidate of $w_2$ in the other language, and $w_2$ has to be the translation candidate of $w_1$. This symmetry guarantees that the translation candidates are relatively reliable and ensures a one-to-one mapping between the context words of the two languages.[1] A side-effect of this definition is that each word in one language can pair up with at most one translation. In this manner, we achieve the desired effect of overwriting previous dimensions involving the same words, under the assumption that the quality of the space will increase over time.

## 4. EVALUATION OF THE BILINGUAL VECTOR SPACE

This section evaluates the quality of the bilingual lexical induction algorithm presented in Section 3. Our analysis focuses on two language pairs: English–German and English–Dutch.[2]

### 4.1 Setup

Like all distributional models, our algorithm requires a corpus for each relevant language. The larger and more comparable these corpora are, the better the results. We use large newspaper corpora: the Twente Nieuws Corpus (TwNC) for Dutch (250 million words) and the Huge German Corpus (HGC) for German (200 million words). For English, we settled for the British National Corpus (BNC, 100 million words). All corpora are tagged and lemmatized. We then define our test vocabulary for each language on the basis of the corpora as the set of all content words that appear more than 4 times per million words. This results in roughly 10,000 target words per language — nouns, verbs and adjectives.

As seed lexicons, we use two freely available online lexicons, namely the English-Dutch lexicon available from `www.freelang.net` (English–Dutch) and the English–German lexicon from `www.dict.cc`. Table I lists some statistics about the behaviour of the lexicons. Note that the German–English lexicons contain a considerably larger number of translation pairs than the English–Dutch lexicons, with much higher coverage but at the same time a much higher degree of ambiguity.

We use two thirds of the lexicons to seed our algorithm and set one third aside for evaluation. If a word in the seed lexicon had several translations, one of these was selected randomly. The results we report are averages over five random samples;

[1] We also experimented with adding only the most frequent symmetric pairs, based on the hypothesis that more evidence leads to more confidence, but did not find any effect apart from slower convergence.

[2] We believe that the results carry over to other language pairs. See Peirsman and Padó [2010] for results on the language pair English–Spanish.

| Language pair | # test words | mean # translations |
|---|---|---|
| Dutch-English | 3951 | 1.51 |
| English-Dutch | 3815 | 1.56 |
| English-German | 8045 | 3.25 |
| German-English | 7339 | 3.56 |

Table I. Statistics of the gold standard lexicons. Two thirds of these lexicons were used for initialization of the space, one third was set aside for testing.

we generally found that this selection does not have a major effect on the results. We also note that we obtained very similar results with much smaller seed lexicons that consisted of automatically identified pairs of cognates and loanwords in another study [Peirsman and Padó 2010]. These findings indicate that the size and quality of the lexicon are not of primary importance, given that the bootstrapping procedure effectively helped filter out incorrect translation pairs and add more newly identified mutual nearest neighbors.

For the computation of monolingual co-occurrences, we use a relatively small context window of three words to each side of the target word in order to focus on taxonomic rather than topical similarity [Turney and Pantel 2010]. We transform raw co-occurrence counts by pointwise mutual information, a standard choice in vector space semantics [Lowe 2001; Chambers and Jurafsky 2009] and calculate distributional similarity by cosine similarity, another standard choice.

## 4.2 Lexicon-based Evaluation

Our first evaluation corresponds to the standard "in-vitro" evaluation procedure for bilingual lexicons, namely a fully automatic comparison against a translation lexicon, in our case the test portion of our lexicons. Let *test* be the test set, $trans_{nn}$ a function that assigns a target language translation to a source language word based on the nearest neighbor function of our bilingual space, and $trans_{gs}$ a function that assigns each source word a *set* of correct translations based on the gold standard lexicon. Furthermore, let *range* denote the range of a function. Then the accuracy of $trans_{nn}$ with respect to $trans_{gs}$ is defined as:

$$acc = \frac{||\{w \in test \,|\, trans_{nn}(w) \in trans_{gs}\}||}{||\{w \in test \,|\, range(trans_{nn}) \cap trans_{gs}(w) \neq \emptyset\}||} \tag{1}$$

The numerator is a straightforward count of correct translations, taking into account only the single nearest neighbor (1-best) in the semantic space. The denominator ensures that test items all correct translations of which are missing in the semantic space are excluded from evaluation, as is standard procedure in the literature [Fung and McKeown 1997].

Figure 2 shows the evolution of the accuracy of the translation candidate throughout the first five iterations of the bootstrapping procedure. As we hypothesized in Section 3.1, we can achieve a considerable improvement in accuracy by adding symmetric translation candidates from the space as new dimensions. Convergence takes place after a small number of iterations, however.

In terms of results, the accuracy of noun translations is generally between 45% and 60%, and the accuracy of verb translations between 35% and 40%. At first glance, this looks like a disappointing result: only about half of the nouns and

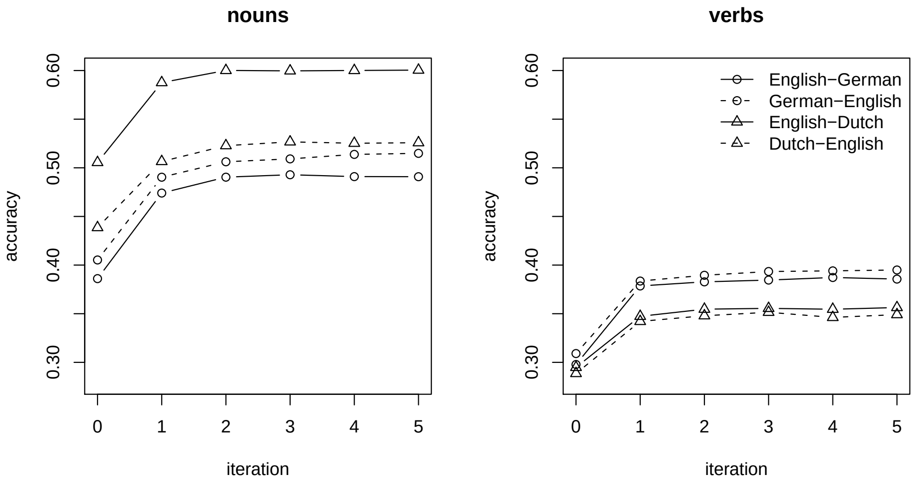

Fig. 2. Accuracy of translation candidates (1-best) at different steps of the bootstrapping process for nouns (left) and verbs (right).

somewhat more than one third of the translation candidates from the semantic space are correct. Although limited coverage of the lexicons is a possible confound, this seems unlikely: The English–Dutch and English–German lexicons differ clearly in this respect (cf. Table I), but the evaluation does not show clear differences: While English–Dutch tends to show better results on nouns, we find better results on verbs for English–German.

However, we believe that it would be premature to conclude at this point that our strategy is a failure. The numbers that we observe are as much a result of our model as a result of the evaluation procedure, namely the comparison against a translation lexicon. Translation lexicons are supposed to contain only instances of translational equivalence proper, i.e., the cross-lingual equivalent of synonymy. Thus, this evaluation discards all instances of non-literal translations as wrong. At the same time, it is well known from monolingual vector spaces that a high degree of semantic similarity can arise not only from synonymy, but also from a range of other semantic relations [Padó and Lapata 2007; Turney and Pantel 2010]. This warrants an analysis of our bilingual spaces in terms of semantic relations.

### 4.3 Relation-based Evaluation

In this section, we assess what semantic relations exist in our bilingual vector spaces. We manually analyze a random sample of 200 nominal and 200 verbal translation pairs for both language pairs. We concentrate again, as in the previous section, on target words and their translation candidates, i.e., their single nearest cross-lingual neighbors.

The basis of our analysis is formed by the standard translation lexicons used by human translators for English–Dutch and English–German [Martin and Tops 2006a; 2006b; Springer 2000; 2003]. We classify translation pairs into nine semantic

| Relation | German | Dutch | Example | Meta-Relation |
|---|---|---|---|---|
| in lexicon | 86 | 99 | Verhältnis ‘relationship’ - relationship | true synonymy (46%/51%) |
| synonym | 5 | 2 | Umstrukturierung ‘restructuring’ - reorganization | |
| antonym | 1 | 3 | Inneres ‘interior’ - exterior | taxonomic similarity (21%/20%) |
| near-synonym | 8 | 2 | Einschätzung ‘estimation’- opinion | |
| hypernym | 15 | 6 | Dramatiker ‘playwright’ - poet | |
| hyponym | 3 | 4 | Kunstwerk ‘work of art’ - painting | |
| co-hyponym | 15 | 24 | Straßenbahn ‘tram’ - bus | |
| related | 39 | 42 | Kapitel ‘chapter’ - essay | relatedness (19%/21%) |
| unrelated | 28 | 18 | DDR-Zeit ‘GDR era’ - trainee | errors (14%/9%) |
| total | 200 | 200 | | |

Table II. Distribution of semantic relations between English noun targets and their translation candidates (1-best) in German and Dutch (examples from German).

relations, under the assumption that these relations can be applied to bilingual word pairs as well. We group the nine relations into four general *meta-relations* (see Table II). The relations are as follows: *In lexicon* means that the translation candidate is listed as a translation in the lexicon. *Synonym*s are translations that should be in the lexicon according to our judgment, but are not. *In lexicons* and *synonyms* together form the meta-relation of *true synonymy*. Next, *antonym*, *near-synonym*, *co-hyponym*, *hyponym* and *hypernym* cover translation candidates that stand in one of these taxonomic, WordNet-inspired relations to a translation in the lexicon. They form the meta-relation *taxonomic similarity*. *Related* describes words and their translation candidates that are semantically associated along other dimensions than taxonomic ones [Budanitsky and Hirst 2006]. Finally, *unrelated* covers translation candidates that are not semantically associated to their target in any way — in other words, outright *errors*. The numbers in the rightmost column of Table II give percentages for the meta-relations for English–German and English–Dutch, respectively.

We find fairly similar patterns for the two language pairs. True synonymy and taxonomic similarity clearly dominate. Together, they account for around 70% of the translation candidates for both target languages. The presence of at least a handful of synonyms that are not in the translation lexicon highlights again the limits of automatic evaluation against translation lexicons, even professional ones. The most frequent taxonomic similarity relation is co-hyponymy. Hypernymy is more frequent than hyponymy, due to the occurrence of compounds in Dutch and German that correspond to multi-word units in English. Still, a considerable number of nearest neighbors are not taxonomically similar to their target — 30% for Dutch, and 33% for German. This relatively high number may be due to the limited comparability between the English corpus on the one side and the Dutch and German corpora on the other. However, most of these translation candidates are still semantically related to their target in some manner. Only a minority — 14% for German and 9% for Dutch — is completely unrelated. For German, a considerable number of these cases stems from adjectives that were misclassified as nouns by the POS tagger.

| Relation | German | Dutch | Example | Meta-Relation |
|---|---|---|---|---|
| in lexicon | 96 | 88 | wassen ‘wash’ - wash | true synonymy (50%/45%) |
| synonym | 4 | 1 | oplaaien ‘erupt’ - erupt | |
| near-synonym | 4 | 3 | profiteren ‘profit’ - benefit | taxonomic similarity (13%/17%) |
| similar full | 15 | 27 | belemmeren ‘hinder’ - facilitate | |
| similar partial | 6 | 4 | gelijkspelen ‘draw’ - beat | |
| related full | 17 | 10 | verslijten ‘wear out’ - mend | relatedness (23%/20%) |
| related partial | 28 | 29 | aanhouden ‘arrest’ - kill | |
| unrelated | 30 | 38 | wemelen ‘bristle with’ - translate | errors (15%/19%) |
| total | 200 | 200 | | |

Table III. Distribution of semantic relations between English verb targets and their translation candidates (1-best) for German and Dutch (Examples from Dutch).

Compared to nouns, semantic relations between verbs are more difficult to classify along taxonomic dimensions [Fellbaum 1998]. For verbs we therefore only distinguish two taxonomic relations: *near-synonymy* and *similarity*. We keep the relation (and meta-relation) *relatedness* for pairs of verbs that are clearly related, but not in a taxonomic way. Typical examples are events and their causes or result states. A second difference to nouns lies in the syntactic behavior of verbs. Two nouns that we classified as semantically related almost always display different syntactic behavior. This is not true for verbs. Non-synonymous verbs often still have valencies that correspond entirely or partially, for example because they belong to the same narrative schema [Chambers and Jurafsky 2009], while synonymous verbs can differ in their valencies. To account for this aspect, we additionally classify the verbal translation candidates with regard to their syntactic correspondence as *full* or *partial*. Full correspondence means the entire argument structure of the word is preserved by its translation candidate: the translation has the same arguments as the source verb, and these arguments fulfill the same roles. This is often true for antonyms (like *open* and *close*), or related verbs (like *meet* and *greet*), which can take the same types of arguments in the same syntactic roles. Table III gives some cross-lingual examples. Partial correspondence means that only some of the syntactic arguments of the source verb are preserved in the translation, either in the same syntactic realization or not. An intra-lingual example of this category is the pair *say* and *talk*, which share the subject argument, although *say* is transitive while *talk* is not.

The resulting distribution is shown in Table III. In contrast to the results in the earlier automatic evaluation, the number of synonymous translation candidates is roughly the same for verbs as for nouns, which indicates that the coverage of available online lexicons is worse for verbs than for nouns. Among non-synonymous translation candidates, we find less taxonomically similar verbs, as compared to nouns; still, around 60% of the translation candidates are taxonomically similar, and 80% are at least related. Not surprisingly, taxonomically similar verbs typically share their complete argument structure with their translation, whereas related verbs do so rarely. At the same time, there is at least partial syntactic overlap in a large majority of cases: only 15% (German) or 19% (Dutch) of the translation candidates are unrelated.

A natural follow-up question is to what extent we can distinguish among these

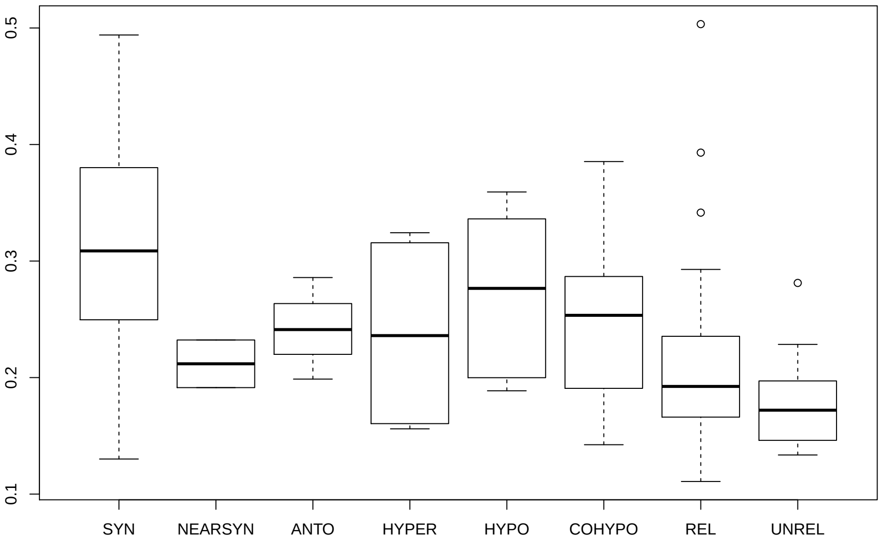

Fig. 3. Cosine values for the semantic relations observed between the Dutch target nouns and their translation candidate (1-best).

relations based on the degree of semantic similarity in the bilingual space. To answer this question, we compute the mean semantic similarity for the pairs of each relation in Tables II and III and compare them with a Kruskal-Wallis test, a nonparametric one-way analysis of variance. We find that both for verbs and nouns, and for both language pairs, the cosine similarities in fact differ significantly between the relations. For illustration, Figure 3 shows a boxplot for the Dutch nouns (results for the other cases look very similar). We clearly see the tendency of similarities to fall for "less close" relations. However, due to the variation in cosine values within the groups, most individual pairwise differences are not significant. In other words, it is not straightforward to classify translation pairs into semantic relations purely on the basis of their semantic similarity.

### 4.4 Discussion

The two evaluations that we have performed leave us with contradictory assessments of the quality of the translation pairs drawn from bilingual vector spaces. According to the translation lexicon-based evaluation, a substantial part of them is simply wrong. Meanwhile, the manual evaluation suggests that a majority of these candidates should be considered partially correct in that they are closely related to the literal translation, most of them even through a well-defined taxonomic relationship. Which one of these evaluations should we believe in?

To answer this question, we can draw an analogy between our bilingual case and the better-researched monolingual case. Monolingually, it is well-established that semantic relations other than strict synonymy can play an important role. Examples include query expansion in IR [Bhogal et al. 2007], smoothing in language models [Dagan et al. 1999], the learning of paraphrases [Lin and Pantel 2001], question answering [Harabagiu et al. 2000] or textual entailment [Dagan et al. 2006]. What these tasks have in common is that they are faced with sparse data, i.e., they need to generalize from seen to new, unseen data. An alternative perspective on this problem is to see it as a "lexical chasm" [Berger et al. 2000] which arises from the ability to linguistically realize the same meaning in a large number of different ways which need to be recognized as equivalent. Obviously, the differences between these realizations go far beyond synonymy, and it is evident that a synonymy-centered evaluation of a semantic space will underestimate the support that the information from the space can provide to semantic processing tasks. We believe that substantially the same situation holds in our bilingual setting.

Of course, this is not to say that using the unfiltered similarity information from a bilingual space is always beneficial: taking other semantic relations into consideration also carries the risk of adding only remotely related or even unrelated terms. This means that in practice, the benefit of doing so depends strongly on two properties of the task at hand. The first one is the degree of *sparsity* – the more sparse the data, the more a task can benefit from other semantic relations. The second one is the degree to which the classes induced by the task are *closed under semantic relatedness*, that is, to what extent semantically related terms are assigned to the same class. If this is not the case, then adding non-synonymy is a bad idea. A clear example of a task whose classes are presumably closed well under relatedness is text classification. A counterexample is the detection of literal translations, where adding non-literal translations would defeat the purpose.

Our agenda for the rest of the paper is to test these hypotheses on two resource induction tasks with opposite profiles concerning the allegedly relevant properties. The first task is the cross-lingual sentiment classification. The classes of positive and negative sentiment triggers are not closed under semantic relatedness: sentiment can be reversed by close taxonomic relations like antonymy (*good* – *bad*), or even near-synonymy (*feast* – *meal* – *grub*). At the same time, for this task, which classifies individual words, the sparse data problem is not serious. We therefore expect that a focus on synonymy will improve results. Our second task is the cross-lingual modeling of verbal selectional preferences — that is, the prediction of the plausibility of predicate-relation-argument triples. Selectional preferences generally apply to fairly broad semantic classes and are therefore closed fairly well under semantic relatedness. At the same time, the prediction of selectional preferences requires co-occurrence statistics for pairs of verbs and arguments, which are quite sparse. Thus, we make the inverse prediction for this task: we expect that the model can benefit from a wider range of semantic relations. For example, verb translations that are related to their target through a narrative schema, and which are classified as "related partial" in Table III, can still provide information about the syntactic and semantic behavior of that target [Chambers and Jurafsky 2009].

A methodological question that we have avoided so far is how to manipulate our bilingual vector space model so as to yield either predominantly synonyms or predominantly non-synonyms. Cut-offs based on semantic similarity are one

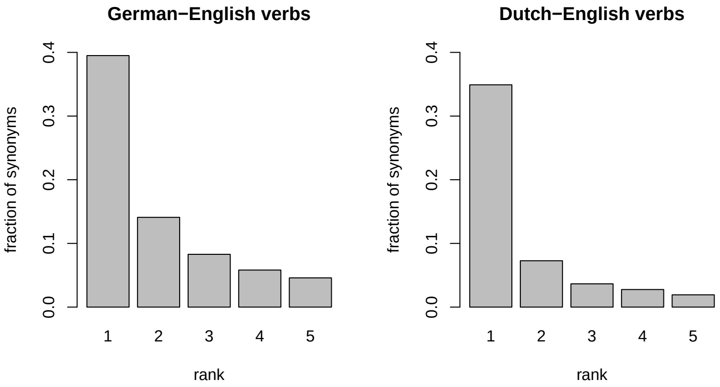

Fig. 4. Fraction of synonyms for the five nearest cross-lingual neighbors for German-English verb translations at iteration five (convergence)

possibility, but they might leave some of the original terms without any translation. In addition, our analysis in Section 4.3 has found that semantic similarity is only imperfectly correlated with semantic relation. For these reasons, we will vary instead the number of nearest neighbors in the vector space that we take into account. Recall that the evaluations in Sections 4.2 and 4.3 only considered the translation candidate for each target, that is, the nearest neighbor. A considerable fraction of these nearest neighbors are synonyms; however if we consider more distant cross-lingual neighbors in the vector space, the proportion of synonyms declines rapidly. Figure 4 shows the traditional translation accuracy of the first through fifth nearest cross-lingual neighbors at iteration five for German-English and Dutch-English verbs (results for nouns are very similar). This behavior suggests that varying the number of neighbors to take into account is a variable with which we can vary the fraction of synonymy effectively.

For space reasons, the task-based evaluation will concentrate on one language pair, namely English–German. For the selectional preference task, a parallel study for English–Spanish can be found in Peirsman and Padó [2010].

## 5. TASK-BASED EVALUATION 1: TRANSFER OF A SENTIMENT LEXICON

### 5.1 Sentiment Analysis and Polarity Transfer

Sentiment analysis is the task of extracting subjective information from texts, like movie or book reviews, and classifying them according to their polarity, that is, as positive or negative [Pang and Lee 2008]. This task has become a field of growing importance in computational linguistics, not least because of its practical relevance. Sentiment analysis is usually approached by (1) determining the positive or negative value of the individual words in the text and (2) combining these into an overall score [Turney 2002; Kim and Hovy 2004; Wilson et al. 2005].

This strategy presupposes the existence of a sentiment lexicon that associates words with their polarity, a classical resource bottleneck. Although a number of studies has addressed the construction of sentiment lexicons in recent years [Wilson et al. 2005; Esuli and Sebastiani 2006; Waltinger 2010; Remus et al. 2010], they are only available for a few languages, and all but the smallest ones are constructed automatically and thus noisy. Not surprisingly, cross-lingual knowledge induction is an attractive perspective for the creation of corresponding resources for lesser-researched languages [Mihalcea et al. 2007; Banea et al. 2008; Scheible et al. 2010].

In this section, we analyze a simple lexicon-based sentiment transfer method based on the bilingual semantic spaces induced in Section 4. In line with the predictions from Section 4.4, we we are interested in the effect of using more than one translation candidate for each target word. Our hypothesis is that sentiment analysis works best when the lexicon contains as few non-synonyms as possible.

### 5.2 Data and Method

This experiment aims at inducing a sentiment lexicon for German words (nouns, verbs and adjectives) on the basis of SentiWordNet, a large sentiment resource for English [Esuli and Sebastiani 2006], which contains a subset of 1000 WordNet synsets manually annotated with sentiment scores. To make the test set as comparable as possible to previous work, we sample 200 words from this manually annotated portion. The German translations of these English words (according to our bilingual space) form our test set.

We experiment with the number of cross-lingual neighbors that participate in the transfer, formally expressed through a function $tr(w, k)$ that maps a word $w$ onto its $k$ nearest cross-lingual neighbors in the bilingual space. For example, $tr(sehen, 3)$ returns the set of the three nearest English neighbors for the German verb *sehen ‘see’*.

We assign polarity scores to our German test set words by averaging the polarity scores of their $n$ nearest cross-lingual (English) neighbors using $tr$. Let score denote a function assigning a polarity score to a word $w$, and subscripts $g$ and $e$ denote German and English, respectively. Then the translated score tr-score$_n$ for a German word using $n$ nearest neighbors (we vary $n$ between 1 and 5) is defined as:

$$\text{tr-score}_n(w_g) = \sum_{w_e \in tr(w_g, n)} \text{score}(w_e)/n \qquad (2)$$

The problem we encounter with this setup is that many of the German test words' English neighbors are outside the manually annotated portion of SentiWordNet and are therefore missing polarity judgments. For this reason, we collect polarity judgments from Amazon's Mechanical Turk crowdsourcing platform [Snow et al. 2006; Mohammad and Turney 2010] for both German and English. Similar to the procedures adopted for the original SentiWordNet annotation as well as by Scheible et al. [2010], we ask native speakers of either English or German to rate the positive and negative connotation of each word on a Likert scale ranging from one (not positive or negative) to five (very positive or negative), independent of each other. Participants were informed that most words were either neutral, positive or negative, but that some words could be both positive and negative, like *joke* or *unexpected*. Each word was judged by five participants, whose polarity scores were averaged.[3]

[3] In order to limit the number of words to be scored on AMT, we only take into account cross-lingual

| | 1-best | 2-best | 3-best | 4-best | 5-best |
|---|---|---|---|---|---|
| positive polarity | **.724** | .721 | .709 | .705 | .704 |
| negative polarity | **.642** | .633 | .640 | .633 | .634 |

Table IV. Spearman correlation between human judgments of the positive/negative polarity score of German words and the average scores for their $n$ English nearest neighbors. All correlations are highly significant (p<.001).

| German target | | first neighbor | | second neighbor | |
|---|---|---|---|---|---|
| Feuer ‘fire’ | neg:2.4 | fire | neg:3.4 | flame | neg:1.4 |
| Hund ‘dog’ | pos:2.4 | dog | pos:2.8 | cat | pos:1.0 |

Table V. Two example target words with their two English nearest neighbors.

We verify the reliability of these scores with Spearman's $\rho$, a non-parametric correlation coefficient appropriate for non-normally distributed data. $\rho$ values range between -1 and 1, with -1 denoting perfect negative correlation, 0 no correlation, and 1 perfect positive correlation. Specifically, we compute the correlation between the judgments of each participant and the mean judgments of the other participants. The mean correlations for our four datasets lie between .6 and .7 (English positive .68, English negative .65, German positive .62, German negative .70). This indicates a good reliability of the scores and is comparable to reliability figures obtained by other studies.

### 5.3 Results and Discussion

We evaluate the success of our polarity transfer for each German test word by correlating the predicted polarity scores (i.e., those computed from the English polarity scores) with the human-provided German polarity scores. Again, we use Spearman's $\rho$. The results are shown in Table IV. The correlations are highly significant throughout, which indicates that lexical translation is by and large a suitable method for polarity transfer, even though it does not take context into account. However, a comparison by the number of nearest neighbors bears out our prediction from Section 4.4, that this task can profit from a narrow focus on synonymy. Using just the translation candidate (1-best) is the best way of transferring the connotation of the target words. Correlation decreases somewhat when more cross-lingual neighbors are used.

The examples in Table V illustrate why this is the case. Both *Feuer* and *fire*, its single nearest neighbor, have a mostly negative connotation, with negativity scores of 2.4 and 3.4, respectively. However, *flame*, the second neighbor, has less strongly negative connotations, with a negativity score of 1.4. Adding this second-best neighbor to the model will thus result in a decrease of the observed correlation. Similarly, our English sentiment judgments show that the word *dog* in English, like German *Hund*, receives a fairly positive evaluation (2.8 in English, 2.4 in German). This is not true for *cat*, the second nearest neighbor to *Hund*, with a positivity score of 1.0. Again, adding this second nearest neighbor will decrease correlation.

neighbors with a cosine of at least 0.30. This reduced the number of necessary annotations from 709 to 252.

To better understand why the degradation from increasing $n$ is small nevertheless, we compute the percentage of word pairs for the different relations in SentiWordNet which preserve the dominant polarity (i.e., which are both predominantly positive, or both predominantly negative). The numbers are 89% for synonyms, 80% for co-hyponyms, and still 74% for hyponyms/hypernyms.[4] Thus, even polarity classes are closed under non-synonymous semantic relations to some degree.

In sum, we find that polarity scores are best transferred from English into German when only translation candidates (single nearest neighbors) are used. However, highly similar non-synonymous neighbors still preserve polarity to a certain extent, and at least in our experiments, the addition of more nearest neighbors leads only to a slight degradation.

## 6. TASK-BASED EVALUATION 2: TRANSFER OF SELECTIONAL PREFERENCES

### 6.1 Modeling Selectional Preferences

The second task is the cross-lingual modeling of selectional preferences. Selectional preferences capture the idea that not all words are equally good arguments to a given verb in a particular argument position. For instance, the English verb *to shoot* generally selects for people as its subject, while its direct object can be people or animals. These preferences play an important role in human sentence processing [McRae et al. 1998], and knowledge about them is helpful for tasks like word sense disambiguation [McCarthy and Carroll 2003] or semantic role labeling [Gildea and Jurafsky 2002].

The modeling of selectional preferences can be phrased as the task of predicting plausibility scores for (PREDICATE, RELATION, ARGUMENT) triples. Virtually all current models start out from a set of seen triples, in the form of a syntactic training corpus, over which they need to generalize in order to cover new, unseen triples. Generalization usually takes place either on the basis of a knowledge source like WordNet [Resnik 1996; Abe and Li 1996; Clark and Weir 2002] or a distributional model acquired from a large syntactically annotated corpus [Erk et al. 2010]. Erk et al. estimate the plausibility of a new argument by calculating its mean distributional similarity to other arguments they have observed in a corpus. Let $(p, r, a)$ stand for a predicate-relation-argument triple, and $Seen_r(p)$ for the set of arguments seen in the corpus in relation $r$ to the predicate $p$. The monolingual model $Pl$ computes the plausibility of the triple $(p, r, a)$ as a weighted average of the vector space similarities between $a$ and all alternative arguments $a'$ seen in the relevant position, with the weight being given by the relative frequency of the alternative argument $a'$:

$$Pl(p, r, a) = \sum_{a' \in Seen_r(p)} \frac{f(a')}{Z(r, p)} \cdot sim(a, a') \tag{3}$$

where $Z(r, p) = \sum_{a' \in Seen_r(p)} f(a')$ is the total frequency of all arguments seen with $p$ in relation $r$.

[4] Since WordNet does not encode our category "semantic relatedness" from Section 4.3, we have no numbers for this relation.

## 6.2 Cross-lingual Transfer of Selectional Preferences

For many languages, neither a sufficiently large corpus nor large knowledge sources are available. Bilingual models aim to solve this problem by mapping plausibility queries from one language onto a language such as English where more resources are available [Agirre et al. 2003].

In this study, we combine the model by Erk et al. with our bilingual vector space to transfer selectional preferences from English into German. Step 1 is to translate German (PREDICATE, RELATION, ARGUMENT) triples $(p_g, r_g, a_g)$ into English triples $(p_e, r_e, a_e)$. In step 2, the plausibility for the English triples is predicted with the monolingual model (Eq. 3). This model can be estimated, for example, from the BNC, one of the few available corpora that are neither genre- nor domain-specific.

The source language predicate $p_g$ and its argument $a_g$, can be translated straightforwardly using the bilingual vector space. Like in Experiment 1, we are interested in the effect of non-synonymous translations and experiment with a varying number $n$ of nearest neighbors. We reuse the notation $tr(w, k)$ to denote the $k$ nearest cross-lingual neighbors of a word $w$. The element of German triples that remains most problematic is the relation $r_g$, which cannot be translated using the bilingual space. We sketch two models that use strategies of increasing sophistication for the cross-lingual transfer of $r_g$. In our first cross-lingual model, $Pl_{\text{XL-1}}$, we simply map the relation to its English equivalent — assuming, in effect, that the syntactic realizations of arguments do not change across languages, and considering the top $k$ nearest cross-lingual neighbors for the German predicate. Our model can be expressed in terms of the monolingual English plausibility model $Pl$ as:

$$Pl_{\text{XL-1}}(p_g, r_g, a_g, k) = \max_{p_e \in tr(p_g, k)} Pl(p_e, r_g, tr(a_g, 1)) \qquad (4)$$

Model XL-1 assumes that German subjects correspond to English subjects and direct objects to direct objects. For prepositional objects, which are governed by language-specific pronouns, this is problematic. We therefore map German prepositional objects at the instance level onto the English prepositional object relation that gives the highest plausibility estimate for the German argument $a_g$.

Note that we maximize the plausibility scores obtained for the different nearest neighbors of the German predicate rather than taking an average. We make this choice to avoid situations where the best translation results in a predicate-argument combination in the target language that is dispreferred for collocational reasons, or where the best translation has a syntactic structure that is not compatible with that of the source verb. For example, both *attend* (first neighbor) and *participate* (second neighbor) are correct translations of the German verb *teilnehmen*. However, only *participate* has the event in a prepositional phrase, like *teilnehmen*. The resulting higher plausibility score for that event suggests that this is the better choice.

Our second model XL-2 is syntax-aware and can deal better with this type of cross-lingual syntactic variation. We first identify the English relation that best corresponds to the source language one by computing the pairwise similarities between the relations of the German verb and all relations observed with the English verb translations, and choosing the English function $r_{opt}$ for which the source language fillers are most plausible:

$$r_{opt}(p_g, r_g, p_e) = \underset{r_e}{\operatorname{argmax}} \sum_{a_g \in Seen_{r_g}(p_g)} Pl(p_e, r_e, tr(a_g, 1)) \quad (5)$$

$Pl_{\text{XL-2}}$ replaces the identity assumption that XL-1 makes for the relation with this mapping:

$$Pl_{\text{XL-2}}(p_g, r_g, a_g, k) = \max_{p_e \in tr(p_g, k)} Pl(p_e, r_{opt}(p_g, r_g, p_e), tr(a_g, 1)) \quad (6)$$

Note that XL-2 requires at least a small German corpus with syntactic analysis to yield $Seen_{r_g}(p_g)$, the set of arguments seen in relation $r_g$ with the predicate $p_g$.

### 6.3 Data and Methods

We evaluate this model on the German plausibility judgment dataset collected by Brockmann [2002]. This dataset contains human judgments of the plausibility of ninety German verb-argument combinations, including subjects, objects, and prepositional objects. We ascertained the reliability of this dataset by re-eliciting human judgments for all triples through Amazon Mechanical Turk, parallel to our first experiment. German native speakers were asked to rate the plausibility of the verb-argument combinations on a five-point scale. We obtained between 7 and 11 judgments for each triple, of which we computed the means. Our own judgments and Brockmann's original data show an almost perfect Spearman $\rho$ correlation of .90. The mean correlation between the judgments of each participant and the mean judgments of the other participants is .77. Both figures speak to the replicability of the data, and can be seen as an upper bound for any modeling approaches.

We predict selectional preferences employing both transfer models, XL-1 and XL-2. The underlying English plausibility model is trained on a version of the BNC parsed with MINIPAR [Lin 1993]. For the cross-lingual transfer, we vary the number of nearest neighbors for the predicate translation between 1 and 5, as in Experiment 1. As the syntactically annotated corpus for XL-2, we use a 30 million word subset of the Huge German Corpus that was parsed with a lexicalized CFG [Schulte im Walde et al. 2001].

For evaluation, we compute the Spearman $\rho$ correlation between the predicted plausibility scores against human judgments. This follows previous work [Resnik 1996; Brockmann and Lapata 2003].

### 6.4 Results and Discussion

Table VI shows the results of our two cross-lingual selectional preference transfer models. The predictions of all of our models are highly significantly correlated with the human judgments. Table VII provides four points of comparison for the "all arguments" condition. The top half shows two simple baselines. The argument frequency baseline uses as its prediction $f(a)$ counted on the parsed portion of the HGC mentioned above, but does not show any correlation with human judgments. The triple frequency baseline uses $f(p, r, a)$ as prediction and already achieves a highly significant correlation, indicating that co-occurrence frequency is a very good proxy of plausibility [Chambers and Jurafsky 2010]. Note however that 44 of the 90 predictions are zero, because the corresponding triples were not seen in the corpus.

| XL-1: Cross-lingual identity of syntactic relations | | | | | |
|---|---|---|---|---|---|
| | 1-best | 2-best | 3-best | 4-best | 5-best |
| subject | .53 | .51 | **.56** | **.56** | .55 |
| direct objects | .58 | .61 | .61 | **.64** | .58 |
| prepositional objects | .33 | .45 | .45 | **.46** | .42 |
| all arguments | .34 | .41 | .44 | **.46** | .40 |

| XL-2: Intelligent syntactic relation mapping | | | | | |
|---|---|---|---|---|---|
| | 1-best | 2-best | 3-best | 4-best | 5-best |
| subjects | .49 | **.63** | .61 | .53 | .48 |
| direct objects | .58 | .66 | .69 | .70 | **.71** |
| prepositional objects | .45 | .50 | .51 | .51 | **.52** |
| all arguments | .34 | .45 | **.46** | .43 | .42 |

Table VI. Spearman correlation between human judgments of German selectional preferences and automatic judgments modeled on the basis of the BNC. All correlations are highly significant (p<.001).

| Point of comparison | Spearman's $\rho$ |
|---|---|
| Argument frequency baseline (on parsed HGC) | -0.13 (n.s.) |
| Triple frequency baseline (on parsed HGC) | 0.30 (p<.01) |
| Corpus-based selectional preference model [Erk et al. 2010], trained on parsed HGC (result from Peirsman and Padó [2010]) | 0.33 (p<.001) |
| Ontology-based selectional preference model [Resnik 1996] (result from Brockmann and Lapata [2003]) | 0.37 (p<.001) |

Table VII. Monolingual German baselines and selectional preference models, "all arguments" condition. In round brackets: Significance levels of the correlations.

The bottom half of the table shows the performance of two monolingual selectional preference models, since there is no previous multilingual work on this dataset. The triple frequency baseline is outperformed by an instantiation of the Erk et al. model trained on the same data, which in turn does worse than the best contender model, one of Brockmann and Lapata's ontology-based models.

When we turn to our cross-lingual models, we see that both XL-1 and XL-2 manage to significantly outperform these monolingual contenders. Unsurprisingly, the best results for XL-2 exceed those for XL-1, indicating that a small amount of syntactic knowledge about the target language improves results. In contrast to the polarity projection task, we observe that the models that use only the translation candidate (1-best) are consistently outperformed by models that use several cross-lingual nearest neighbors. This is consistent with our main hypothesis.

For XL-1, we find a clear optimum of using four translation candidates, where we obtain the best results for all types of relations. For this model the information provided by additional neighbors outweighs the potential noise introduced by their non-synonymy. Although the details are different, a similar trend is visible for XL-2. Again, the model that uses only the best translation candidate is always outperformed. The estimated plausibilities for the direct and prepositional objects reach an optimum at five nearest neighbors, while those for the subject relation peak at two nearest neighbors. We attribute this to the less severe sparsity problem for subjects: on average, the set of seen arguments for subjects, $Seen_{subj}$, is twice as large as the number of seen arguments for direct objects, $Seen_{obj}$, and even larger when compared to prepositional objects. Thus, direct and prepositional objects benefit more from reduced sparsity.

In sum, models of selectional preferences profit consistently from the inclusion of additional nearest neighbors in the cross-lingual transfer model, even if these are predominantly non-synonymous with the target word. A closer look at actual instances supports the interpretation, in line with our hypotheses in Section 4.4. Specifically, semantically similar (i.e., taxonomically related) words (cf. Table III), like hyponyms, hypernyms or antonyms, refer to similar events with similar participants, and thus typically have (almost) identical selectional preferences as the target word. To return to our previous example, if *open* is translated by a verb that means *close*, this incorrect translation is still informative about the selectional preferences of the target word. In our data, the German triple *(Luft, obj, reinigen)* '(air, obj, clean)' (the 9th most plausible object triple in the data) is initially translated into English correctly, but the triple *(air, obj, clean)* receives a relatively low plausibility (rank 19 of 30 direct object triples). The second-nearest neighbor to *reinigen*, however, is its antonym *pollute*, which is judged more likely to have *air* as its object (rank 11). This rank is much closer to the German one, and leads to an improved prediction.

The situation is somewhat more subtle for semantically related verbs that do not stand in a taxonomic relationship, since they do not necessarily describe similar events. However, many of those verb pairs form *narrative chains*, i.e., likely sequences of events [Chambers and Jurafsky 2009] that share one or more participants. For example, if one of the neighbors of *meet* is the associated event *grüßen 'greet'*, this can still contribute informative arguments for the relations of *meet* (groups of people, personal names, etc.). An example from our data is German *stagnieren 'stagnate'* which is paired with English *boost*. These verbs can be conceptualized as forming part of an "economical development" narrative chain, where periods of stagnation, growth, and recession alternate. Again, it seems plausible that the subject of *stagnieren* and the object of *boost* realize the same participant and share arguments. Consequently, even the "wrong" translation *boost* for *stagnieren* is informative with respect to the selectional preferences of the German verb.

## 7. CONCLUSIONS AND OUTLOOK

This article has presented an analysis of the semantic relations in a bilingual lexicon extracted from a bilingual vector space. The space was constructed on the basis of independent monolingual corpora, using a bootstrapping process that is initialized with a small seed translation lexicon and iteratively adds newly acquired translations as dimensions. Such seed lexicons are available for many language pairs and can be acquired for others via bridge languages [Mausam et al. 2009]. Also, the ability of the bootstrapping process to correct errors in the lexicon makes it possible to use even lists of cognates and loanwords as the initial lexicon [Peirsman and Padó 2010].

The standard method for evaluating automatically acquired bilingual lexicons is a comparison against human-created translation lexicons, which corresponds to an exclusive focus on (cross-lingual) synonymy. Such an evaluation ignores other semantic relations such as taxonomic similarity and semantic relatedness, which make up a significant fraction of unfiltered translation candidate pairs in the bilingual lexicon. We have argued that monolingual and cross-lingual semantic processing tasks share the "lexical chasm" as their main challenge. Cross-lingual tasks are therefore able to profit from knowledge about semantic relations and semantic similarity in the same way as monolingual tasks. However, the practical benefit depends to two factors: (a), the magnitude of the sparsity problems; (b), the extent to which similar and related words share the semantic behavior of interest. In two task-based evaluations, we have found that for polarity transfer, a task with low sparsity and highly lexically specific behavior, the inclusion of non-synonymy relations leads to a small, but manageable, degradation. For the transfer of selectional preferences, a task with high sparsity where the classes of interest are closed under semantic similarity, we see a clear improvement for the inclusion of non-synonymous, related or similar translation candidates.

Our conclusion from these results is that there is no single best way to build bilingual lexicons for different cross-lingual NLP tasks. Instead, the first step in the induction of bilingual lexicons for cross-lingual applications should be an analysis of the positive or negative impact of different lexical-semantic relations on the application. We suggest that our clustering of relations into the meta-relations "synonymy" – "taxonomic similarity" – "semantic relatedness" can serve to build *profiles* for applications. Polarity transfer profits only from the first group, true synonymy, while selectional preference transfer can profit from all three, at least with regard to verbs. Conceivably, textual inference-related tasks might profit from synonymy and taxonomic similarity, but not from mere relatedness, at least not without further qualifications.

This perspective of course raises the question of how to filter out individual meta-relations in practice. Our analysis found that cosine similarity in our bilingual space is presently merely an imperfect predictor of semantic relations. In future work, we want to extend monolingual strategies to optimize semantic spaces for particular relations to the bilingual case, for example through the use of asymmetrical similarity measures which promise to be better indicators of taxonomic relationships [Michelbacher et al. 2007; Kotlerman et al. 2009].

In the present study, we have only considered two language pairs (English–German and English–Dutch), and performed the task-based evaluation only on the first one. The three languages are also fairly closely related. This is a clear limitation of our study, and it must be expected that bilingual spaces for less related languages may be of lower quality, containing more weakly related or unrelated translation pairs. Nevertheless, there are some indications that linguistic distance is not the only important influence on translation quality. First, in our manual evaluation, English–Dutch and English–German outperformed each other on nouns and verbs, respectively, although Dutch is more closely related to English. Second, in Peirsman and Padó [2010] we investigated the projection of selection preferences for English–Spanish, obtaining results that were almost as good as for English–German. We hypothesize that there is an interaction between linguistic distance and the properties of the task, as discussed above, and that the two factors must be seen in conjunction: Tasks that are more closed under similarity and relatedness will tend to suffer less from the noise introduced by a larger linguistic distance, and vice versa.

## REFERENCES

Abe, N. and Li, H. 1996. Learning word association norms using tree cut pair models. In *Proceedings of the 13th International Conference on Machine Learning*. Bari, Italy, 3–11.

Agirre, E., Aldezabal, I., and Pociello, E. 2003. A pilot study of English selectional preferences and their cross-lingual compatibility with Basque. In *Proceedings of the 6th International Conference on Text, Speech and Dialogue*. České Budějovice, Czech Republic, 12–19.

Banea, C., Mihalcea, R., and Wiebe, J. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of the 6th Language Resources and Evaluation Conference*. Marrakech, Morocco.

Baroni, M. and Lenci, A. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics 36,* 4, 673–721.

Berger, A., Caruana, R., Cohn, D., Freitag, D., and Mittal, V. 2000. Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. 192–199.

Bhogal, J., Macfarlane, A., and Smith, P. 2007. A review of ontology based query expansion. *Information Processing & Management 43,* 4, 866–886.

Brockmann, C. 2002. Evaluating and combining approaches to selectional preference acquisition. M.S. thesis, Universität des Saarlandes, Saarbrücken, Germany.

Brockmann, C. and Lapata, M. 2003. Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary, 27–34.

Budanitsky, A. and Hirst, G. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics 32,* 1, 13–47.

Chambers, N. and Jurafsky, D. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. Singapore, 302–610.

Chambers, N. and Jurafsky, D. 2010. Improving the use of pseudo-words for evaluating selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, 445–453.

Chiao, Y.-C. and Zweigenbaum, P. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*. Taipei, Taiwan, 1208–1212.

Clark, S. and Weir, D. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics 28,* 2, 187–206.

Dagan, I., Glickman, O., and Magnini, B. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges*, J. Quinonero-Candela, I. Dagan, B. Magnini, and F. d'Alché Buc, Eds. Lecture Notes in Computer Science, vol. 3944. Springer, 177–190.

Dagan, I., Lee, L., and Pereira, F. C. N. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning 34,* 1-3, 43–69.

De Smet, W. and Moens, M.-F. 2009. Cross-language linking of news stories on the web using interlingual topic modelling. In *Proceedings of the CIKM Workshop on Social Web Search and Mining*. Hong Kong, 57–64.

Diab, M. and Resnik, P. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, 255–262.

Dumais, S., Landauer, T., and Littman, M. 1996. Automatic cross-linguistic information retrieval using Latent Semantic Indexing. In *Proceedings of the SIGIR Workshop on Cross-Linguistic Information Retrieval*. Zurich, Switzerland, 16–23.

Erk, K., Padó, S., and Padó, U. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics 36,* 4, 723–763.

Esuli, A. and Sebastiani, F. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of 5th Language Resources and Evaluation Conference*. Genoa, Italy.

Fellbaum, C., Ed. 1998. *WordNet. An Electronic Lexical Database.* The MIT Press, Cambridge, MA.

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

Fung, P. and Yee, L. Y. 1998. An IR approach for translating new words from non-parallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*. Montreal, Canada, 414–420.

Gamallo Otero, P. 2008. Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. In *Proceedings of the LREC-2008 Workshop on Comparable Corpora*. Marrakech, Morocco, 19–26.

Garera, N., Callison-Burch, C., and Yarowsky, D. 2009. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Proceedings of the 13th Conference on Natural Language Learning*. Boulder, CO, 129–137.

Gildea, D. and Jurafsky, D. 2002. Automatic labeling of semantic roles. *Computational Linguistics 28,* 3, 245–288.

Haghighi, A., Liang, P., Kirkpatrick, T. B., and Klein, D. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. Columbus, OH, 771–779.

Harabagiu, S. M., Moldovan, D. I., Pasca, M., Mihalcea, R., Surdeanu, M., Bunescu, R. C., Girju, R., Rus, V., and Morarescu, P. 2000. Falcon: Boosting knowledge for answer engines. In *Proceedings of the Text Retrieval Conference*. Gaithersburg, MD.

Harris, Z. 1954. Distributional structure. *Word 10,* 2/3, 146–162.

Holmlund, J., Sahlgren, M., and Karlgren, J. 2005. Creating bilingual lexica using reference wordlists for alignment of monolingual semantic vector spaces. In *Proceedings of the 15th Nordic Conference on Computational Linguistics*. Joensuu, Finland, 71–77.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering 11,* 3, 311–325.

Kilgarriff, A. and Yallop, C. 2000. What's in a thesaurus? In *Proceedings of the 2nd Language Resources and Evaluation Conference*. Athens, Greece, 1371–1379.

Kim, S.-M. and Hovy, E. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland, 1367–1373.

Kotlerman, L., Dagan, I., Szpektor, I., and Zhitomirsky-Geffet, M. 2009. Directional distributional similarity for lexical expansion. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. Singapore, 69–72.

Landauer, T. K. and Dumais, S. T. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review 104,* 2, 211–240.

Lee, L. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, MA, 25–32.

Lin, D. 1993. Principle-based parsing without overgeneration. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Columbus, OH, 112–120.

Lin, D. and Pantel, P. 2001. Discovery of inference rules for question answering. *Natural Language Engineering 7,* 4, 342–360.

Lowe, W. 2001. Towards a theory of semantic space. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*. Edinburgh, UK, 576–581.

Markó, K., Schulz, S., Medelyan, O., and Hahn, U. 2005. Bootstrapping dictionaries for cross-language information retrieval. In *Proceedings of the 28th Annual International ACM*

*SIGIR Conference on Research and Development in Information Retrieval.* Salvador, Brazil, 528–535.

Martin, W. and Tops, G. A. J., Eds. 2006a. *Van Dale Groot Woordenboek Engels-Nederlands (Electronic edition 2.1).* Utrecht/Antwerp: Van Dale Lexicografie.

Martin, W. and Tops, G. A. J., Eds. 2006b. *Van Dale Groot Woordenboek Nederlands-Engels (Electronic edition 2.1).* Utrecht/Antwerp: Van Dale Lexicografie.

Mausam, Soderland, S., Etzioni, O., Weld, D., Skinner, M., and Bilmes, J. 2009. Compiling a massive, multilingual dictionary via probabilistic inference. In *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.* Association for Computational Linguistics, Suntec, Singapore, 262–270.

McCarthy, D. and Carroll, J. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics 29,* 4, 639–654.

McRae, K., Spivey-Knowlton, M., and Tanenhaus, M. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language 38,* 3, 283–312.

Merlo, P., Stevenson, S., Tsang, V., and Allaria, G. 2002. A multilingual paradigm for automatic verb classification. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* Philadelphia, Pennsylvania, 207–214.

Michelbacher, L., Evert, S., and Schütze, H. 2007. Asymmetric association measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing.* Borovets, Bulgaria.

Mihalcea, R., Banea, C., and Wiebe, J. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.* Prague, Czech Republic, 976–983.

Mohammad, S., Gurevych, I., Hirst, G., and Zesch, T. 2007. Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* Prague, Czech Republic, 571–580.

Mohammad, S. and Turney, P. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.* Los Angeles, CA, 26–34.

Och, F. J. and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics 29,* 1, 19–52.

Padó, S. and Lapata, M. 2007. Dependency-based construction of semantic space models. *Computational Linguistics 33,* 2, 161–199.

Pang, B. and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2*, 1–135.

Peirsman, Y. and Geeraerts, D. 2009. Predicting strong associations on the basis of corpus data. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.* Athens, Greece, 648–656.

Peirsman, Y., Heylen, K., and Geeraerts, D. 2008. Size matters. Tight and loose context definitions in English word space models. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics.* Hamburg, Germany, 9–16.

Peirsman, Y. and Padó, S. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Proceedings of the 11th Meeting of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies.* Los Angeles, CA, 921–929.

Rapp, R. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics.* Cambridge, MA, 320–322.

Rapp, R. 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.* College Park, MA, 519–526.

Remus, R., Quasthoff, U., and Heyer, G. 2010. SentiWS — a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th Language Resources and Evaluation Conference*. Valletta, Malta.

Resnik, P. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition 61*, 127–159.

Riloff, E. and Shepherd, J. 1999. A corpus-based bootstrapping algorithm for semi-automated semantic lexicon construction. *Journal of Natural Language Engineering 5,* 2, 147–156.

Sadat, F., Yoshikawa, M., and Uemura, S. 2003. Learning bilingual translations from comparable corpora to cross-language information retrieval: hybrid statistics-based and linguistics-based approach. In *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages*. Sapporo, Japan, 57–64.

Sahlgren, M. 2006. The Word-Space model. Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-dimensional Vector Spaces. Ph.D. thesis, Stockholm University, Stockholm, Sweden.

Scheible, C., Laws, F., Michelbacher, L., and Schütze, H. 2010. Sentiment translation through multi-edge graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China, 1104–1112.

Schulte im Walde, S., Schmid, H., Rooth, M., Riezler, S., and Prescher, D. 2001. Statistical grammar models and lexicon acquisition. In *Linguistic Form and its Computation*, C. Rohrer, A. Rossdeutscher, and H. Kamp, Eds. CSLI Publications, Stanford, CA, 389–440.

Snow, R., Jurafsky, D., and Ng, A. Y. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia, 801–808.

Springer, O., Ed. 2000. *Langenscheidts Enzyklopädisches Wörterbuch der englischen und deutschen Sprache ("Der Große Muret-Sanders") Englisch-Deutsch*, 12. ed. Langenscheidt.

Springer, O., Ed. 2003. *Langenscheidts Enzyklopädisches Wörterbuch der englischen und deutschen Sprache ("Der Große Muret-Sanders") Deutsch-Englisch*, 9. ed. Langenscheidt.

Turney, P. and Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research 37*, 141–188.

Turney, P. D. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. 417–424.

Waltinger, U. 2010. GermanPolarityCues: A lexical resource for German sentiment analysis. In *Proceedings of the 7th Language Resources and Evaluation Conference*. electronic proceedings, Valletta, Malta.

Wilson, T., Wiebe, J., and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, BC, Canada, 347–354.

Yarowsky, D. and Ngai, G. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*. Pittsburgh, Pennsylvania, 200–207.

Zeman, D. and Resnik, P. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*. Hyderabad, India, 35–42.

Zhao, H., Song, Y., Kit, C., and Zhou, G. 2009. Cross language dependency parsing using a bilingual lexicon. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. Singapore, 55–63.