Construction of a Benchmark Data Set for Cross-lingual Word Sense Disambiguation

Els Lefever[1,2] and Véronique Hoste[1,2]

[1]LT3, Language and Translation Technology Team, University College Ghent
Groot-Brittanniëlaan 45, 9000 Gent, Belgium
els.lefever, veronique.hoste@hogent.be
[2]Department of Applied Mathematics and Computer Science, Ghent University
Krijgslaan 281 (S9), 9000 Gent, Belgium

Abstract

Given the recent trend to evaluate the performance of word sense disambiguation systems in a more application-oriented set-up, we report on the construction of a multilingual benchmark data set for cross-lingual word sense disambiguation. The data set was created for a lexical sample of 25 English nouns, for which translations were retrieved in 5 languages, namely Dutch, German, French, Italian and Spanish. The corpus underlying the sense inventory was the parallel data set Europarl. The gold standard sense inventory was based on the automatic word alignments of the parallel corpus, which were manually verified. The resulting word alignments were used to perform a manual clustering of the translations over all languages in the parallel corpus. The inventory then served as input for the annotators of the sentences, who were asked to provide a maximum of three contextually relevant translations per language for a given focus word. The data set was released in the framework of the SemEval-2010 competition.

1. Introduction

Word Sense Disambiguation (WSD), the NLP task of selecting the correct sense of an ambiguous word in a given context, is a well-researched problem (see for example Agirre and Edmonds (2006) and Navigli (2009)), which still suffers from some shortcomings. First of all, the creation of large sense inventories and sense-tagged corpora is both time-consuming and expensive, and as a result such sense inventories and corpora are very scarce for languages other than English. Furthermore, there is a growing feeling in the WSD community that WSD should not be considered as a stand-alone problem but should be integrated and evaluated in real applications such as Machine Translation.

In this paper, we describe the construction of a multilingual benchmark data set in which the senses are solely based on a parallel corpus. Working with a parallel corpus instead of manually sense-annotated data might reduce the problems described above. Moreover, the use of corpus evidence might be more reliable than human annotations, given the low inter-annotator agreements on sense tagging experiments and the often arbitrary division of word meanings into distinct dictionary senses and entries (Atkins, 1991). Furthermore, using corpus translations instead of human-defined sense labels should facilitate the integration of a dedicated WSD module in multilingual applications. The methodology to deduce word senses from parallel corpora is based on the hypothesis that different meanings of a polysemous word can be lexicalized across languages. Such an approach also implicitly deals with the problem of sense granularity, as finer sense distinctions are only relevant as far as they get lexicalized in the retrieved translations. Many WSD studies have already shown the validity of this cross-lingual evidence idea (Gale et al., 1993; Ide et al., 2002; Ng et al., 2003; Apidianaki, 2009). However, given the lack of multilingual data sets which can serve as benchmark data for the evaluation of cross-lingual word sense disambiguation systems, it remains unclear how viable multilingual WSD is, and it is difficult to assess how well different approaches perform on this task.
In the framework of the SemEval-2010 (Evaluation Exercises on Semantic Evaluation) competition[1], we formulated a Cross-lingual Word Sense Disambiguation task (Lefever and Hoste, 2009) for which we developed (i) a sense inventory in which the sense distinctions were extracted from a multilingual corpus; and (ii) a lexical sample data set in which the ambiguous words were annotated with the senses from the multilingual sense inventory. Both resources were constructed for a lexical sample of 25 nouns. The data set was divided into a trial set of 5 ambiguous nouns and a test set of 20 nouns.

The remainder of this article is organized as follows. Section 2 focuses on the construction of the sense inventory, describing both the word alignments and the manual clustering of the aligned words. In Section 3, we discuss the annotation of the benchmark data set with the senses from the multilingual sense inventory. Finally Section 4 concludes this paper.

2. Construction of the sense inventory

The document collection which serves as the basis for the gold standard sense inventory is the Europarl parallel corpus[2], which is extracted from the proceedings of the European Parliament (Koehn, 2005). We selected 6 languages from the 11 European languages represented in the corpus, viz. English (our target language), Dutch, French, German, Italian and Spanish. All data were already sentence-aligned using a tool based on the Gale and Church algorithm (Gale and Church, 1991), which was part of the Europarl corpus.

[1] http://semeval2.fbk.eu/semeval2.php
[2] http://www.statmt.org/europarl/

We only considered the 1-1 sentence alignments between English and the five other languages[3] (see also (Tufiş et al., 2004) for a similar strategy). After the selection of all English sentences containing the target nouns and the aligned translations in the five target languages, the following two steps were taken for both the trial and test data in order to obtain a multilingual sense inventory.

1. word alignment on the sentences to find the set of possible translations for the set of ambiguous nouns, and manual evaluation of the word alignments (Section 2.1.)
2. manual clustering by meaning (per target word) of the resulting translations (Section 2.2.)

This multilingual sense inventory was then used for the manual annotation of new instances containing the ambiguous lexical sample words. We return to this in Section 3.

2.1. Word Alignment

In order to detect the possible translations for the set of ambiguous nouns, we performed word alignment on the selected Europarl sentences by means of GIZA++ (Och and Ney, 2003). An example of these word alignments (marked in bold) for the word *occupation* is given in the sentences below.

(1) *SOURCE*: This monitoring committee , which is part of OLAF , comprises five independent experts who continue to pursue their normal **occupations** however .
SPANISH: Este comité supervisor de la OLAF está formado por cinco expertos independientes pero que realizan su **actividad profesional** normal .
DUTCH: Het Comité van toezicht van OLAF bestaat uit onafhankelijke experts , die ook hun normale **beroepsbezigheden** blijven voortzetten.
GERMAN: Dieser Überwachungsausschuss des OLAF besteht aus fünf unabhängigen Experten , die aber ihrer normalen **Berufstätigkeit** nachgehen .
FRENCH: Ce comité de suivi de l' OLAF est composé de cinq experts indépendants qui poursuivent cependant leur **activité professionnelle** normale .
ITALIAN: Detto comitato è composto da cinque esperti indipendenti , che tuttavia non cessano di svolgere la loro **attività lavorativa** precedente .

As example (1) clearly illustrates, one single target word can lead to multiword translations, such as for example *actividad profesional* in Spanish; and to compounds, such as *beroepsbezigheden* in Dutch and *Berufstätigkeit* in German, which concatenate the parts of compounds in one orthographic unit. In both cases, we kept the multipart translation as a valid translation suggestion.

All GIZA++ alignment links were manually verified by certified translators in all six languages. The human annotators were instructed to correct wrong word alignment and assign a “NULL” link to words for which no valid translation could be identified.

While checking the word alignment output, the annotators were also asked to provide extra information in a dedicated remarks section for the four specific remark categories illustrated below:

1. the translation is a *compound* that corresponds to an English multiword

(2) *SOURCE*: Firstly , the promotion of and participation in the drafting of joint plans for the creation of an integrated services network in the transport and energy sectors , with the backing of the European **Investment Bank** .
DUTCH: In eerste instantie moet men in gemeenschappelijk overleg met de Europese **Investeringsbank** , en met haar steun , projecten uitwerken en uitvoeren voor de totstandkoming van geïntegreerde dienstennetwerken in de vervoer- en energiesector .

35-11	Bank	Investeringsbank
Remarks:	compound	Investment Bank

2. there is a *fuzzy link* between the target word and its translation

(3) *SOURCE*: The situation of Palestine is extremely serious due to the **occupation** of its land by Israel (...)
DUTCH: De Palestijnen verkeren in een bijzonder moeilijke situatie aangezien hun vaderland door Israël **wordt bezet** (...)

10-18	occupation	wordt bezet
Remarks:	fuzzy	(the occupation of: wordt bezet)

3. there is a *tokenisation* problem (e.g. the English target word is part of a hyphenated compound whereas only the target word itself should be considered)

(4) *SOURCE*: The suspicions raised by the Court of Auditors are fully confirmed by the **non-movements** at international level (...)
FRENCH: l' absence de **flux** internationaux confirme pleinement les soupçons de la Cour des comptes (...)

10-18	movement	flux
Remarks:	tokenisation	(non-movements)

4. the target word is used with a different part-of-speech tag (other than *noun*) and therefore marked as *wrong input*, meaning that it should not be considered for building up the sense inventory

(5) *SOURCE*: (...) the entire directive would be undermined by those who choose to **bank** in certain preferential tax havens.
DUTCH: (...) zou de gehele richtlijn ondergraven worden door degenen die hun **geld onderbrengen** in bepaalde bevoorrechte belastingsparadijzen .

10-18	bank	geld onderbrengen
Remarks:	wrong input	(PoS)

[3] This six language sentence-aligned subsection of Europarl can be downloaded at `http://lt3.hogent.be/semeval`.

	Dutch	French	Spanish	Italian	German
bank (total: 4029 instances)					
Compound	31%	3.4%	0.7%	11%	73%
Fuzzy link	0.6%	4.4%	0.8%	0.8%	1.4%
Tokenisation	1.7%	1.8%	0.1%	0.1%	1.3%
Wrong Input	0.3%	3.1%	0.3%	0.3%	2.3%
movement (total: 4221 instances)					
Compound	20.9%	2.1%	1%	4.1%	66%
Fuzzy link	4.1%	2.6%	1.7%	1%	4.1%
Tokenisation	0.3%	0.6%	0%	0.4%	0.6%
Wrong Input	0%	0%	0%	0%	0%
occupation (total: 633 instances)					
Compound	8.5%	0.3%	0%	8%	10.1%
Fuzzy link	9.1%	7.6%	0.5%	1.6%	6.6%
Tokenisation	1.6%	30.6%	0.2%	1.7%	2.4%
Wrong Input	0%	0%	0%	0%	0%
passage (total: 237 instances)					
Compound	3.4%	0.8%	0.4%	9.7%	1.3%
Fuzzy link	19%	19.8%	11.8%	3.4%	18.1%
Tokenisation	0%	2.5%	0%	0%	0%
Wrong Input	0%	0%	0%	0%	0%
plant (total: 1631 instances)					
Compound	46.8%	7.9%	7.8%	22.4%	54.8%
Fuzzy link	2.6%	8.3%	1.3%	1%	4.5%
Tokenisation	0.5%	1.1%	0%	1.3%	0.8%
Wrong Input	1.3%	2.6%	1.5%	1.3%	2%

Table 1: Percentages of remark categories per word in the trial data

Table 1 gives an overview of the frequency of all remark categories (expressed in percentages of the total amount of instances) for the trial data. As expected, compound translations tend to occur very frequently in German and Dutch (up to 73% for German), and much less frequently in the romance languages. Another intuition that appears to be confirmed by these figures is that more abstract words (such as *passage*) are generally more freely translated, resulting in a higher percentage of fuzzy links between the two corresponding translation units.

The considerable proportion of compound translations also results in a higher number of different translations for Dutch and German, which has important consequences for the multilingual WSD task the data set has been designed for. In multilingual WSD systems, the sense label typically consists of a translation, whereas in more traditional WSD approaches, the label consists of a sense that is picked from a predefined sense inventory (such as WordNet (Fellbaum, 1998)). As a consequence the multilingual WSD systems for Dutch and German will have a broader set of classes (or translations) to choose from, which makes the WSD task more complicated. Figure 1 illustrates this by listing the number of different translations (or classes in the context of WSD) for all trial and test words.

2.1.1. Word alignment performance

In a next step we calculated the performance of the automatically generated word alignments against our manually validated word alignment reference. A straightforward measure for word alignment performance is the F-score, which combines precision and recall. The following formulas were used to calculate precision, recall and F-score on all word-to-word links for our target words, with R referring to the reference set of manually generated alignments and A referring to the automatic alignments generated by the system:

$$Precision = \frac{|A \cap R|}{|A|} \quad (1)$$

$$Recall = \frac{|A \cap R|}{|R|} \quad (2)$$

$$F{-}score = \frac{2 * Precison * Recall}{Precision + Recall} \quad (3)$$

For the trial words, we calculated precision, recall and F-score on all five language pairs (with English as the source language, and the other five languages as target languages), covering 10,751 sentence pairs in total. The average precision was 90.0%, the average recall 85.2% and the average F-score 87.9%. Strikingly, we observed considerable differences in alignment performance between the trial words, with F-scores ranging from as low as 70.2% for *passage* to as high as 89.7% for *bank*. In general we see that word alignment performance seems to be related to the degree in which a word is abstract in meaning, suggesting that more abstract words such as *passage* are more challenging for

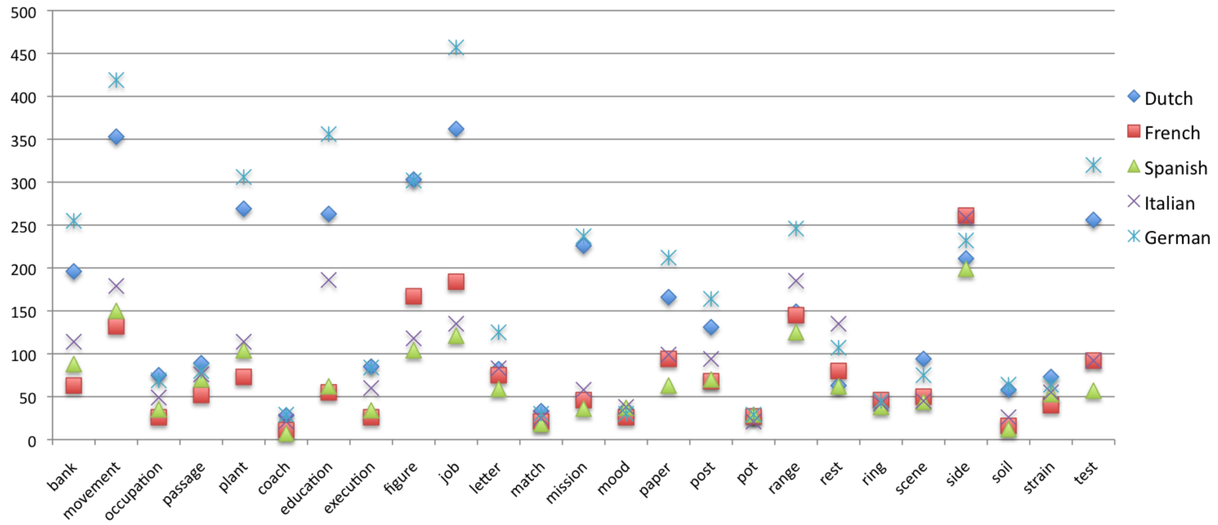

Figure 1: Number of different translations per word for Dutch, French, Spanish, Italian and German.

word alignment. This would appear to be due to the fact that these words are generally translated more freely. They are often paraphrased[4], or have more or less free translational correspondences[5], typically resulting in fuzzy or even "NULL" links.

2.1.2. Inter-annotator agreement

Additionally, the inter-annotator agreement was checked on a sample of 6,530 instances. Again, we used formulas (1), (2) and (3) to calculate precision, recall and F-score, with R now referring to the set of word-to-word alignments of the first annotator and A referring to the set of alignments that were verified by the second annotator. The overall inter-annotator agreement is very high (with an F-score of 95.1%), but once again we observe a lower performance for the more abstract words such as *passage* (F-score of 81%). It turns out that for these words, the annotators are often unsure whether to consider free translations as fuzzy links, or rather to assign the "NULL" link label.

2.2. Manual Clustering

After the manual verification, the resulting translations were clustered per meaning by one annotator. In order to do so, the translations were coupled across languages on the basis of the unique sentence IDs. After the selection of all unique translation combinations, the translations were grouped into clusters. The clusters were organized in two levels, in which the top level reflects the main sense categories (e.g. for the word *bank* we have (1) financial meaning, (2) supply or stock, (3) sloping land beside water, (4) West Bank and (5) group of similar objects), and the subclusters represent the finer sense distinctions. Translations that correspond to English multiword units were identified and in case of non-apparent compounds (e.g. not marked with a hyphen), the different compound parts were separated by §§ in the clustering file. All clustered translations were also manually lemmatized. As an example, Table 2 gives an overview of the different clusters that were created for the test word *coach*.

3. Annotation of trial and test instances

The resulting sense inventory was used to annotate the sentences in the trial (20 sentences per ambiguous word) and test set (50 sentences per ambiguous word), which were all extracted from the JRC-ACQUIS Multilingual Parallel Corpus[6] and the British National Corpus[7]. In total, 1100 sentences were annotated. For the annotation of the ambiguous target words in the sentences, we proceeded in the following way: the annotators were asked to (a) pick the contextually appropriate sense cluster and to (b) choose their three preferred translations from this cluster. In case they could not find three appropriate translations, they were also allowed to provide fewer. These potentially different translations were used to assign frequency weights (as shown in Table 5) to the gold standard translations per sentence. Example (6) below shows the annotation result in both French and Italian for an English source sentence containing *bank*. The part of the sense inventory they used for labeling this sentence is displayed in Table 3.

(6) SENTENCE 3. Considering the importance of the existing links between the Community and the Palestinian people of the West **Bank** and the Gaza Strip, (...)

French Cluster: 4

French 1: Cisjordanie
French 2: rive
French 3: bande

Italian Cluster: 4

Italian 1: Cisgiordania
Italian 2: riva
Italian 3: sponda

[4] E.g. the English "waiting for *passage* to" can be translated in Dutch as "wachten op *toestemming om te reizen* naar", which literally means "waiting for *the permission to travel* to".

[5] E.g. the English "*passage* of time" can be translated in French as "*jour après jour*", which literally means "*day after day*".

[6] `http://wt.jrc.it/lt/Acquis/`

[7] `http://www.natcorp.ox.ac.uk/`

	Dutch	**German**	**French**	**Spanish**	**Italian**
1. sports manager/handler					
	coach	Trainer	entraîneur	entrenador	allenatore
	speler-trainer	Coach	capitaine		
	trainer	National§§trainer			
football coach	voetbal§§trainer	Fußbal§§trainer	entraîneur	entrenador	allenatore
2. bus/autobus					
	streekbus	Reise§§bus	autocar	autocar	autobus
	autobus	Bus§§transport	car	autobús	corriera
	bus	Linien§§bus	bus		autocorriera
	toerbus	Bus§§verkehr	autobus		pullman
	touringcar	Omnibus			autobus di linea
		Bus			pulmino
		Omnibus§§dienst			a mezzo pullman
		Kraft§§omnibus			corriere di frequente
		Reisen§§bus			trasporto in autobus
					autocarro
					mezzi pesanti
Coach Directives	bus§§richtlijn	Bus§§richtlinie	bus	autocar	autobus
	touringcar	Busverkehrs§§unternehmer	autocar		
coach driver	bus§§dienst	Bus§§reise	autocar	autocar	autobus da turismo
					pullman
coach company	bus§§maatschappij	Bus§§unternehmen	autocar	autocar	pullman
	bus§§onderneming	Bus§§unternehmer			operatore del trasporto
	touringcar§§bedrijf	Omnibus§§unternehmen			settore
					autocorriera
coach service	touringcar	Bus§§unternehmer	autocar	autocar	autobus
	touringcar§§dienst	Linien§§verkehrs§§dienst			trasporto in autobus
coach journey	bus§§chauffeur	Bus§§fahrer	autocar	autocar	autobus
coach crash	bus§§ongeluk	Bus§§unfall	autocar	autocar	corriera
coach passenger	bus§§passagier	Bus§§reisende	autocar	autocar	autobus
coach travel	bus§§reis	Busreise§§anbieter	car	autobús	
coach transport	bus§§toerisme	Bus§§tourismus	autobus	autocar	corriera
	bus§§vervoer	Bus§§reise	car		pullman
			autocar		
coach trip	schoolreisje			autocar	
coach operator	touringcar§§operator	Bus§§unternehmer	autocariste	autocar	autobus
3. Carriage **3.1 General**					
	koets	Post§§kutsche	diligence	diligencia	diligenza
3.2 Drive coach and horses					
	als een olifant in een porseleinkast tekeer gaan	gröblichst mißachten	battre en brêche	saltarse a la torera	buttare all'aria
	de effecten tenietdoen van	ad absurdum führen	remettre en question	dar al traste con	minare l'esistenza stessa
4. passenger car, part of train					
	trein§§wagon	Waggon	wagon	vagón	treno
	wagon				vagone

Table 2: Translation cluster for the English noun *coach* in the test set

Table 4 illustrates the agreement on the appropriate sense cluster for the five trial words. The first two columns represent the average number of clusters and top clusters per sentence, and the annotator consensus scores can be read from the last two columns. The agreement scores simply represent the number of sentences (out of 20) for which all annotators agree on the cluster (column 4) or on the top cluster (column 5). The results show that there is fairly little consensus when also incorporating the subclusters, but they also show a clear cluster consensus on the less abstract words, which is directly reflected in the number of sentences on which all annotators agree.

For each instance, the gold standard that results from the manual annotation contains a cluster number and a set of

Dutch	Italian	French	German	Spanish
Bank	Cisgiordania	Cisjordanie	West§§jordanland	Cisjordania
Cisjordanië	sponda	rive	Jordan§§ufer	río Jordán
Jordaan-oever	cisgiordano	bande	West-Bank	Franja
Jordaan§§oever	riva occidentale del Giordano	cisjordanien	West§§bank	costa
Transjordanië	sponda occidentale del Giordano	Bank	Bank	orilla
West§§bank	Bank	Banque	West§§jordangebiet	Ribera
West§§oever	striscia di Gaza		West§§jordanien	junto
deel	riva		West §§jordan§§ufer	
oever			West§§küste	
			West§§ufer	
			Ufer	

Table 3: Translation cluster for the English noun *bank* in the *West Bank* meaning

Target word	Avg Nr cl	Avg Nr top cl	Sent w/ cl cons	Sent w/ top cl cons
Bank	1.15	1.05	18	19
Plant	2	1.05	5	19
Passage	2	1.25	9	16
Occupation	2.15	1.7	5	9
Movement	2.9	1.7	1	9

Table 4: Overview of the annotator consensus for the 5 words in the trial data

translations that are enriched with frequency information. The format of both the input file and gold standard is similar to the format that will be used for the SemEval Cross-Lingual Lexical Substitution task (Sinha et al., 2009). Table 5 lists the six-language gold standard for the trial sentence in example (6):

Language	gold standard translations and frequency weights
French	bande 2; cisjordanie 5; cisjordanien 1; rive 3;
Dutch	cisjordanië 1; jordaanoever 3; oever 2; westbank 3; westoever 3;
Italian	cisgiordania 3; riva 1; riva occidentale del giordano 2; sponda 1; sponda occidentale del giordano 1; striscia di gaza 1;
Spanish	cisjordania 4; franja 3; ribera 1; río jordán 2;
German	west-bank 1; westbank 2; westjordanien 2; westjordanland 2; westjordanufer 3; westufer 2;

Table 5: Gold standard for the target word *bank* for the trial sentence in example (6)

4. Conclusion

We described the construction of a multilingual benchmark data set, which was developed in the framework of the SemEval-2010 Cross-lingual Word Sense Disambiguation task (Lefever and Hoste, 2009). On the basis of a multilingual sense inventory, which was induced from the Europarl parallel corpus, we annotated a set of English sentences with their corresponding translations in five languages.
In future work, we will use this data set to develop and test a cross-lingual approach to word sense disambiguation.

5. References

E. Agirre and P. Edmonds, editors. 2006. *Word Sense Disambiguation*. Text, Speech and Language Technology. Springer, Dordrecht.

M. Apidianaki. 2009. Data-driven semantic analysis for multilingual wsd and lexical selection in translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Athens, Greece.

B. Atkins. 1991. Building a lexicon: the contribution of lexicography. *International Journal of Lexicography*, 3:167–204.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Computational Linguistics*, pages 177–184.

W.A. Gale, K.W. Church, and D. Yarowsky. 1993. A method for disambiguating word senses in a large corpus. In *Computers and the Humanities*, volume 26, pages 415–439.

N. Ide, T. Erjavec, and D. Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit*.

E. Lefever and V. Hoste. 2009. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the NAACL-HLT 2009 Workshop: SEW-2009 - Semantic Evaluations*, pages 82–87, Boulder, Colorado.

R. Navigli. 2009. Word sense disambiguation: a survey. In *ACM Computing Surveys*, volume 41, pages 1–69.

H.T. Ng, B. Wang, and Y.S. Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 455–462, Santa Cruz.

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

R. D. Sinha, D. McCarthy, and R. Mihalcea. 2009. Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the NAACL-HLT 2009 Workshop: SEW-2009 - Semantic Evaluations*, Boulder, Colorado.

Dan Tufiş, Radu Ion, and Nancy Ide. 2004. Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 1312–1318, Geneva, Switzerland, August. Association for Computational Linguistics.