An Approach to Acquire Word Translations from Non-Parallel Texts*

Pablo Gamallo Otero[1] and José Ramom Pichel Campos[2]

[1] Departamento de Língua Espanhola, Faculdade de Filologia
Universidade de Santiago de Compostela, Galiza, Spain
pablogam@usc.es

[2] Departamento de Tecnologia Linguística da Imaxin|Software
Santiago de Compostela, Galiza
jramompichel@imaxin.com

Abstract. Few approaches to extract word translations from non-parallel texts have been proposed so far. Researchers have not been encouraged to work on this topic because extracting information from non-parallel corpora is a difficult task producing poor results. Whereas for parallel texts, word translation extraction can reach about 99%, the accuracy for non-parallel texts has been around 72% up to now. The current approach, which relies on the previous extraction of bilingual pairs of lexico-syntactic templates from parallel corpora, makes a significant improvement to about 89% of words translations identified correctly.

1 Introduction

In the last decade, many works have been carried out to automatically extract word and/or multi-word translations from bilingual parallel corpora [15, 1, 20, 13]. These works share a common strategy: they perform first the alignment of segments and then, on the basis of such an alignment, they compute word correspondences in each pair of segments. In some of these experiences, word-level translation accuracy achieved very high scores: about 99%. Unfortunately, the amount of available bilingual parallel texts is still small, specially in specific, academic or technological domains. Given a particular knowledge domain, the number of parallel texts is much lower than that of monolingual texts. This type of texts are known as comparable, *non-parallel* corpora. Nowadays, in the World Wide Web, comparable non-parallel corpora are more prevalent than parallel corpora. However, the highest rate to date in word-level translation from non-parallel corpora is relatively small, 72% [18], in comparison to the accuracy rate achieved from parallel corpora.

* This work has been supported by Ministerio de Educacin y Ciencia of Spain, within the project GARI-COTERM, ref: HUM2004-05658-D02-02.

In this paper, we present a method to extract word translations, which takes advantage from the positive aspects of both parallel (high accuracy) and non-parallel corpora (high coverage). The strategy we propose is the following. We first extract, from small parallel texts, a representative set of bilingual correspondences between unambiguous lexico-syntactic templates. Then, these pairs of bilingual templates are used as *local contexts* to extract word translations from comparable, non-parallel corpora. A word $w1$ in the source language can be the translation of a word $w2$ in the target language if $w1$ tends to occur in local contexts that are translations of the local contexts $w2$ occurs in. Using this method, we achieved 89% accuracy, which is close to the scores reached by the extraction approaches from aligned, parallel texts. In the following sections, we first recall previous approaches to word translations from non-parallel text, then we describe our own method, and finally, we provide and discuss two different experimental results.

2 Related Work

There are few approaches to extract bilingual lexicons from non-parallel corpora in comparison to those using a strategy based on aligned, parallel texts. The most efficient method to extract word translations from comparable, non-parallel corpora is described in [6, 7, 18]. The starting point of this strategy is as follows: word w_1 is a possible translation of w_2 if the words with which w_1 co-occurs within a particular window are translations of the words with which w_2 co-occurs within the same window. This strategy relies on a list of bilingual word pairs (*seed words*) provided by an electronic bilingual dictionary. w_1 is a candidate translation of w_2 if they tend to co-occur with the same seed words. There are three drawbacks to this method. First, it is not a knowledge-poor approach since it needs external lexical resources such as a bilingual dictionary. Second, not all the words of the dictionary seem to be reliable seed words. Polysemic words should be removed from the list of seed words since they may introduce semantic noise. And third, according to the Harris's hypothesis [12], counting co-occurrences within a window of size N is less precise than counting co-occurrences within local syntactic contexts. In the most efficient approaches to thesaurus generation [11, 14], word similarity is computed using co-occurrences between words and specific syntactic contexts. Syntactic contexts are less ambiguous and more sense-discriminative than contexts defined as windows of size N. In order to overcome these drawbacks, we will use as seed expressions, not word entries from an external bilingual dictionary, but pairs of bilingual lexico-syntactic templates which were extracted from a small parallel corpus. As the lexico-syntactic templates represent unambiguous local contexts of words, they are discriminative and confident seed expressions to extract word translations from non-parallel texts.

There exist other approaches to bilingual lexicon extraction which do not use external bilingual dictionaries [5, 17, 4]. Yet, [5] failed to reach an acceptable accuracy rate for actual use, [17] had strong computational limitations, and [4] was applied only to non-parallel texts in the same language. On the other hand, [3] describes a particular strategy based on a multilingual thesaurus instead of a external bilingual dictionary.

Finally, some researchers have focused on a different issue: disambiguation of candidate translation. According to [16], the process of building bilingual lexicons from non-parallel corpora is a too difficult and ambitious objective. He preferred to work on a less ambitious task: to choose between several translation alternatives previously selected from a bilingual dictionary.

In this paper, the objective of our extraction method is to build bilingual lexicons from comparable non-parallel corpora, with the help of some lexico-syntactic information extracted from small parallel corpora. No external lexical ressource will be used.

3 The Approach

Our method can be divided into the following steps:

- text processing,
- extraction of bilingual lexico-syntactic templates from parallel corpora,
- extraction of word translations from non-parallel texts.

3.1 Text Processing

Parallel and non parallel corpora are tokenized, lemmatized, and tagged. English texts are tagged using TreeTagger [19], while the Spanish part is tagged with Freeling [2]. As no manual correction is made, the extraction of word translations will inherit errors and problems introduced by the tagger. Finally, we identify binary dependencies in order to extract lexico-syntactic templates. From each binary dependency, we extract two complementary lexico-syntactic templates. Table 1 shows some representative examples. A lexico-syntactic template represents the local context of a set of syntactically related words. Given the binary dependency:

of (import, sugar)

template < import of [NOUN] > represents the set of nouns that can appear after “import of”, for instance, “sugar”, “goods”, “oil”, etc. On the other hand, < [NOUN] of sugar > represents the set of nouns appearing before “of sugar”: “import”, “export”, “sell”, etc. We follow the notion of *co-requirement* introduced in [9].

Note that *lobj* represents the relationship between a verb and the noun immediately appearing at its left; *robj* is the relationship between a verb and the noun appearing at its right. On the other hand, *modAdj* is the relationship between a noun and its adjective modifier and *modN* is the relation between two nouns: the head and its modifier.

Binary Dependencies	Templates
of (import, sugar)	< import of [NOUN] > < [NOUN] of sugar >
robj (approve, law)	< approve [NOUN] > < [VERB] law >
lobj (approve, president)	< president, [VERB] > < [NOUN] approve >
modAdj(legal, document)	< legal [NOUN] > < [ADJ] document >
modN(area, protection)	< protection [NOUN] > < [NOUN] area >

Table 1. Binary dependencies and their corresponding templates.

3.2 Extracting Bilingual Templates from Parallel Corpora

Once the lexico-syntactic templates have been identified, we aim at extracting bilingual correspondences between templates from aligned, small parallel corpora. For this purpose, we use a standard learning method, already described and evaluated in [8]. The pairwise correlation between two lexico-syntactic templates is computed by taking into account their occurrence in each aligned segment. We use the Dice coefficient to compute similarity between templates. Each template of the source language is linked to the most similar template of the target language provided that the Dice coefficient is higher than an empirically set threshold. Tables 2, 3, and 4, depict some bilingual correlations extracted from an English-Spanish parallel corpus.

English	**Spanish**	**Dice**
< active [NOUN] >	< [NOUN] activo >	0,65
< african[NOUN] >	< [NOUN] africano >	0,65
< agricultural [NOUN] >	< [NOUN] agrícola >	0,63
< alarming [NOUN] >	< [NOUN] alarmante >	0,45
< albanian [NOUN] >	< [NOUN] albanés >	0,61
< alcoholoic[NOUN] >	< [NOUN] alcohólico >	0,48
< alternative[NOUN] >	< [NOUN] alternativa >	0,75
< ambitious [NOUN] >	< [NOUN] ambicioso >	0,44
< american [NOUN] >	< [NOUN] americano >	0,48
< annual [NOUN] >	< [NOUN] anual >	0,77

Table 2. Bilingual correlations between templates derived from dependencies "noun-adjective" (excerpt of a list extracted from the Proceedings of the European Parliament - EuroParl).

Bilingual links between lexico-syntactic templates will be used as *seed* local contexts in the following step. Such templates will serve as the textual anchor points in non-parallel corpora.

3.3 Extracting a Bilingual Lexicon from Non-Parallel Corpora

We propose the following algorithm for finding word translation pairs from non-parallel corpora:

English	Spanish	Dice
< [NOUN] after year>	< [NOUN] tras año>	0,41
< crime against [NOUN] >	< crimen contra [NOUN] >	0,66
< fight against [NOUN] >	< lucha contra [NOUN] >	0,43
< violence against [NOUN] >	< violencia contra [NOUN] >	0,81
< [NOUN] against discrimination >	< [NOUN] contra discriminación >	0,50
< [NOUN] against fraude >	< [NOUN] contra fraude >	0,48
< [NOUN] against humanity >	< [NOUN] contra humanidad >	1.00
< [NOUN] against poverty >	< [NOUN] contra pobreza >	0,60
< [NOUN] against racialism >	< [NOUN] contra racismo >	0,68
< [NOUN] against woman >	< [NOUN] contra mujer >	0,79

Table 3. Bilingual correlations between templates derived from nominal dependencies with preposition (excerpt of a list extracted from EuroParl).

English	Spanish	Dice
< africa [VERB] >	< frica [VERB] >	0,66
< agreement [VERB] >	< acuerdo [VERB] >	0,45
< agriculture [VERB] >	< agricultura [VERB] >	0,50
< aid [VERB] >	< ayuda [VERB] >	0,57
< alcohol [VERB] >	< alcohol [VERB] >	0,76
< ambition [VERB] >	< ambición [VERB] >	0,58
< amsterdam [VERB] >	< amsterdam [VERB] >	0,63
< angola [VERB] >	< angola [VERB] >	0,70
< animal [VERB] >	< animal [VERB] >	0,44
< article [VERB] >	< artículo [VERB] >	0,50
< assistant [VERB] >	< asistente [VERB] >	0,66

Table 4. Bilingual correlations between templates derived from noun-verb dependencies, i.e. from left-objects (excerpt of a list extracted from EuroParl).

```
Given a bilingual list of lexico-syntactic templates
(i.e., seed contexts):

   - for every word $w_{i}$ in the source language, find the number
     of times it instantiates each template of the bilingual list
     in order to build a vector with this information;
     AND
   - for every word $w_{j}$ in the target language, find the number
     of times it instantiates each template of the bilingual list
     in order to build a vector with this information:

         Compute Weighted Jaccard similarity: $WJ(w_{i},w_{j})$;
         if  $w_{j}$ is the most similar word to $w_{i}$, then
         both words define a translation pair.
```

Let's take an example. Table 5 shows some positions in the feature vector that defines word "president". Each feature corresponds to a lexico-syntactic template extracted from the parallel corpus. The value of the feature is the number of times the word instantiates that template in the non-parallel corpus, i.e., the co-occurrence frequency between the word and the linguistic context associated to the template. Every template in the vector must have a Spanish correlation. The first column of Table 5 represents the position of a feature in the vector.

president		
00034	<conference of [NOUN]>	323
00176	<[NOUN] of council>	218
00182	<[NOUN] of court>	69
00234	<[NOUN] of republic>	35
00701	<former [NOUN]>	69
00776	<new [NOUN]>	63
01324	<[NOUN] declare>	52
01543	<congratulate [NOUN]>	13

Table 5. Excerpt of the vector associated to "president".

Table 6 depicts the corresponding vector positions associated to the Spanish noun "presidente". The features appearing in these positions are then the translations of those depicted in Table 5

presidente		
00034	<conferencia de [NOUN]>	509
00176	<[NOUN] de consejo>	1013
00182	<[NOUN] of tribunal>	54
00234	<[NOUN] of república>	134
00701	<antiguo [NOUN]>	36
00776	<nuevo [NOUN]>	86
01324	<[NOUN] declarar>	90
01543	<felicitar a [NOUN]>	31

Table 6. Excerpt of the vector associated to the Spanish word "presidente".

As tables 5 and 6 shows, the Spanish noun "presidente" occurs in templates that are translations of English templates also instantiated by "president". To compute the degree of similarity between these two words, we use a weighted version of Jaccard coefficient. This measure was proposed by [11] and [10] to generate clusters of semantically related words from monolingual corpora. As in our work, these authors perform a shallow syntactical analysis before constructing the co-occurrence vectors.

Jaccard measure calculates the similarity value between two words by comparing the templates they share and do not share. The weighted Jaccard coefficient considers a global and a local weight for each template . The global weight

gw takes into account how many different words are associated with a given template. It is computed by the following formula:

$$gw(template_j) = 1 + \frac{\sum_i p_{ij} \log_2(p_{ij})}{Ndeps}$$

where

$$p_{ij} = \frac{frequency\, of\, template_j\, with\, word_i}{total\, number\, of\, templates\, of\, word_i}$$

and $Ndeps$ is the total number of binary dependencies extracted from the non-parallel corpus. The local weight lw is the number of times a word instantiates a template:

$$lw(word_i, template_j) = \log_2(frequency\, of\, template_j\, with\, word_i)$$

The whole weight w of a template given a word is the multiplication of both the global and the local weights. So, the weighted Jaccard similarity WJ between two words m and n is computed by:

$$WJ(word_m, word_n) = \frac{\sum_j \min(w(word_m, template_j), w(word_n, template_j))}{\sum_j \max(w(word_m, template_j), w(word_n, template_j))}$$

Each bilingual pair of words is compared by computing WJ similarity. The most similar Spanish words to a given English word are considered as their possible translations.

4 Experiments and Evaluation

4.1 Text Corpora

In our experiments, we used two small English-Spanish parallel corpora: one that was selected from the proceedings of the European Parliament (*EuroPal*); each part contains about 1 million words. The other corpus is constituted by the English and Spanish version of the European Constitution (*EuroConst*). Each part contains about 150,000 words. These two parallel corpora allowed us to extract translation correspondences between lexico-syntactic templates. We learned 2,551 translation pairs from EuroPal, and 1,667 from EuroConst. The reunion of the two sets gives rise to 3,830 different translation pairs.

Word translations were learnt from a larger English-Spanish non-parallel corpus selected from the proceedings of EuroPal. The English part consists of 14 million words while the size of the Spanish part is about 17 million words. The two parts were selected in such a way that no English subparts were translations of Spanish subparts. The bilingual templates we have extracted from the parallel corpus are used here as local contexts of words.

4.2 Results

By selecting only those translation pairs whose Jaccard coefficient is higher than a given threshold (>= 0.1), we generated a lexicon of 11,701 word entries. These entries are lemmas belonging to three syntactic categories: nouns, verbs, and adjectives.

4.3 Evaluation 1: Accuracy

The first evaluation test was inspired from the protocol defined in [18]. 100 test nouns with a frequency of 100 or higher were randomly selected from the English corpus[1]. Using the similarity coefficient, the test nouns was compared to all nouns occurring in the Spanish corpus. Table 7 shows the results for 40 out of the 100 English test nouns. For each English test word, the top five translations are listed. The word considered to be the most expected, salient, and appropriate translation is in bold. As in [18], only one word is selected as the most appropriate translation, except in those cases where the test word can be translated by several nouns considered as synonyms in almost all contexts: for instance, "convention" is correctly translated by both "convenio" and "convención". If we look at Table 7, we see that in most cases our system predicts the most appropriate translation, with other semantically related nouns immediately following. Note that in some cases, behind the most appropriate translation, the other 4 candidates can also be possible translations of the test word: see for instance the candidate translations for "comment", "conference", or "consequence". Even if the first candidate suggested by the system is clearly the most expected translation ("comentario", "conferencia", "consecuencia"), the rest of candidates can be considered as synonyms in certain linguistic contexts.

To measure the accuracy of the system, we counted the number of times where a correct translation of the source word is ranked first. This was true for 89 out of 100 test words (89% accuracy). For comparison, [18] reports an accuracy of 72% and [6] of 30%, when only the top candidate is counted. In another test, we checked whether a correct translation appeared among the top 5 of the ranked list. This was true in 96 cases, while in [18], it was true in 89. This evaluation shows that our system makes a significant improvement in comparison to the two most similar approaches to identification of word translations from non-parallel texts.

4.4 Evaluation 2: Precision and Recall

The previous evaluation does not allow to check the completeness of the translations (recall and precision) since it does not observe the behaviour of word tokens in real text. For this purpose, we selected at random a test corpus containing 150 English word tokens, tagged as nouns, verbs, and adjectives. Each selected word

[1] In [18], the 100 test words were not randomly selected. They belong to a list of stimulus words used in [21] for an association experiment.

English test nouns	**Top 5 candidate translations**				
coherence	credibilidad	**coherencia**	claridad	rigor	visibilidad
cohesion	**cohesión**	integración	solidaridad	estabilidad	igualdad
colleague	**colega**	diputado	amigo	ministro	pueblo
comment	**comentario**	observación	pregunta	reflexión	palabra
commissioner	**comisario**	presidente	colega	diputado	gobierno
commitment	**compromiso**	obligación	respuesta	responsabilidad	criterio
committee	grupo	consejo	parlamento	**comité**	comisión
communication	**comunicación**	información	documento	diálogo	iniciativa
community	unión	**comunidad**	estado	política	país
company	**empresa**	industria	mercado	economía	sector
competition	**competencia**	igualdad	inversión	cooperación	democracia
compromise	consenso	planteamiento	**compromiso**	solución	discusión
concept	**concepto**	idea	planteamiento	enfoque	exigencia
concern	**preocupación**	interés	responsabilidad	exigencia	consideración
conclusion	**conclusión**	resultado	declaración	idea	punto
condition	**condition**	criterio	norma	regla	aspecto
conference	**conferencia**	reunión	cumbre	convención	declaración
confidence	**confianza**	consenso	solidaridad	credibilidad	respaldo
conflict	**conflicto**	crisis	tensión	guerra	diálogo
connection	vínculo	**conexión**	contacto	consulta	vinculación
consequence	**consecuencia**	efecto	repercusión	resultado	impacto
consideration	**consideración**	aspecto	razón	criterio	dimensión
consumer	**consumidor**	población	ciudadano	democracia	empresa
content	**contenido**	alcance	elemento	calidad	orientación
context	**contexto**	marco	vista	plano	término
contribution	**contribución**	aportación	respuesta	papel	apoyo
control	**control**	gestión	responsabilidad	competencia	información
convention	**convenio**	tratado	**convención**	protocolo	carta
cooperation	**cooperación**	relación	colaboración	diálogo	acuerdo
coordination	**coordinación**	colaboración	diálogo	cooperación	equilibrio
cost	**coste**	gasto	necesidad	efecto	consecuencia
country	**país**	estado	región	europa	unión
course	**curso**	año	resultado	plazo	enfoque
court	**tribunal**	ley	autoridad	comité	gobierno
creation	**creación**	integración	existencia	establecimiento	crecimiento
crime	delito	**crimen**	terrorismo	violencia	enfermedad
crisis	**crisis**	conflicto	catástrofe	dificultad	situación
criterion	**criterio**	regla	requisito	condición	exigencia
criticism	**crítica**	comentario	mensaje	petición	observación
culture	**cultura**	sociedad	dimensión	democracia	agricultura
currency	democracia	**moneda**	economía	mercado	sociedad

Table 7. Results for 40 out of 100 test words.

$\mathbf{Prec}_1$	$\mathbf{Prec}_{10}$	$\mathbf{Recall}_1$	$\mathbf{Recall}_{10}$	**Number of bilingual templates**
.72	.88	.63	.77	$2,551$
.74	.92	.66	.81	$3,830$

Table 8. Results concerning precision and recall

was extracted with its immediate context in order to allow evaluators to make appropriate decisions. Word tokens can belong to word types with a frequency of 1 or higher than 1. Given a word token of the test corpus, the evaluators had to decide if there is a correct Spanish translation between the top 10 most similar words. The results are summarized in Table 8. In general terms, we call *Precision* the number of correct translations proposed by the system divided by the number of all translations which have been suggested. *Recall* is the number of correct translations proposed by the system divided by the number of all test instances. Concerning $precision_1$ and $recall_1$ the correct translation is the first word proposed by the system. As regards $precision_{10}$ and $recall_{10}$, the correct translation is within the top 10 candidate translations proposed by the system. In addition, we made two different experiences: in the first experiment, we use as seed contexts the 2,551 bilingual templates extracted from only one parallel corpus: EuroPal. However, in the second experiment, we put together the 3,830 templates extracted from both EuroPal and EuroConst. As Table 8 shows, the higher the number of bilingual templates taken as seed contexts, the better the scores associated to precision and recall. It means that we could improve the learning method if the experiences are made using a larger set of discriminative template pairs. The results of this evaluation cannot be compared to other approaches, since as far as we know this type of experience has not been made yet in related work.

5 Conclusion and Discussion

Few approaches to extract word translations from non-parallel texts have been proposed so far. This is a difficult task. Whereas for parallel texts, word translation extraction reaches in some studies about 90%, the accuracy for non-parallel texts has been around 72% up to now. The main contribution of the approach proposed in this paper is to use bilingual pairs of lexico-syntactic templates as seed expressions. This makes a significant improvement to about 89% of words translations identified correctly.

The ability to identify word translations from non-parallel corpora could be seen as an indicator in favor of the idea that machines, like humans, are able to acquire new linguistic knowledge without being directly exposed to that knowledge.

References

1. Lars Ahrenberg, Mikael Andersson, and Magnus Merkel. A simple hybrid aligner for generating lexical correspondences in parallel texts. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 29–35, Montreal, 1998.
2. X. Carreras, I. Chao, L. Padró, and M. Padró. An open-source suite of language analyzers. In *4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 2004.
3. H. Dejean, E. Gaussier, and F. Sadat. Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora. In *COLING 2002*, Tapei, Taiwan, 2002.
4. Mona Diab and Steve Finch. A statistical word-level translation model for comparable corpora. In *Proceedings of the Conference on Content-Based Multimedia Information Access (RIAO)*, 2001.
5. Pascale Fung. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *14th Annual Meeting of Very Large Corpora*, pages 173–183, Boston, Massachusettes, 1995.
6. Pascale Fung and Kathleen McKeown. Finding terminology translation frmo non-parallel corpora. In *5th Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong, 1997.
7. Pascale Fung and Lo Yuen Yee. An ir approach for translating new words from nonparallel, comparable texts. In *Coling'98*, pages 414–420, Montreal, Canada, 1998.
8. Pablo Gamallo. Extraction of translation equivalents from parallel corpora using sense-sensitive contexts. In *10th Conference of the European Association on Machine Translation (EAMT'05)*, pages 97–102, Budapest, Hungary, 2005.
9. Pablo Gamallo, Alexandre Agustini, and Gabriel Lopes. Clustering syntactic positions with similar syntactic requirements. *Computational Linguistics*, 31(1), 2005.
10. Pablo Gamallo, Caroline Gasperin, Alexandre Agustini, and Gabriel P. Lopes. Syntactic-based methods for measuring word similarity. In V. Mautner, R. Moucek, and K. Moucek, editors, *Text, Speech, and Discourse (TSD-2001)*, pages 116–125. Berlin:Springer Verlag, 2001.
11. Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers, USA, 1994.
12. Z. Harris. Distributional structure. In J.J. Katz, editor, *The Philosophy of Linguistics*, pages 26–47. New York: Oxford University Press, 1985.
13. Oi Yee Kwong, Benjamin K. Tsou, and Tom B. Lai. Alignment and extraction of bilingual legal terminology from context profiles. *Terminology*, 10(1):81–99, 2004.
14. Dekang Lin. Automatic retrieval and clustering of similar words. In *COLING-ACL'98*, Montreal, 1998.
15. Dan Melamed. A word-to-word model of translational equivalence. In *35th Conference of the Association of Computational Linguistics (ACL'97)*, Madrid, Spain, 1997.
16. Hiroshi Nakagawa. Disambiguation of single noun translations extracted from bilingual comparable corpora. *Terminology*, 7(1):63–83, 2001.
17. Reinhard Rapp. Identifying word translations in non-parallel texts. In *33rd Conference of the ACL'95*, pages 320–322, 1995.
18. Reinhard Rapp. Automatic identification of word translations from unrelated english and german corpora. In *ACL'99*, pages 519–526, 1999.

19. Helmut Schimd. Treetagger. In *A language independent part-of-speech tagger*, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html, 2002.
20. Jorg Tiedemann. Extraction of translation equivalents from parallel corpora. In *11th Nordic Conference of Computational Linguistics*, Copenhagen, Denmark, 1998.
21. Manfred Wettler and Reinhard Rapp. Computation of word associations based on the co-occurrences of words in large corpora. In *1st Workshop on Very Large Corpora*, pages 84–93, Columbus, Ohio, 1993.