Bilingual Lexicon Extraction at the Morpheme Level Using Distributional Analysis

Amir HAZEM and Béatrice DAILLE
LINA, Université de Nantes.
2 rue de la houssinière 44000 Nantes, France.
amir.hazem@univ-nantes.fr, beatrice.daille@univ-nantes.fr

Abstract

Bilingual lexicon extraction from comparable corpora is usually based on distributional methods when dealing with single word terms (SWT). These methods often treat SWT as single tokens without considering their compositional property. However, many SWT are compositional (composed of roots and affixes) and this information, if taken into account, can be very useful to match translational pairs, especially for infrequent terms where distributional methods often fail. For instance, the English compound *xenograft* which is composed of the root *xeno* and the lexeme *graft* can be translated into French compositionally by aligning each of its elements (*xeno* with *xéno* and *graft* with *greffe*) resulting in the translation: *xénogreffe*. In this paper, we experiment several distributional modellings at the morpheme level that we apply to perform compositional translation to a subset of French and English compounds. We show promising results using distributional analysis at the root and affix levels. We also show that the adapted approach significantly improve bilingual lexicon extraction from comparable corpora compared to the approach at the word level.

Keywords: Bilingual lexicon extraction, comparable corpora, morphemes, compound term, distributional analysis

1. Introduction

Nowadays comparable corpora are widely used in many applications of natural language processing, particularly in bilingual terminology extraction where parallel corpora are a scarce resource (Rapp, 1995; Fung, 1995; Chiao and Zweigenbaum, 2002; Laroche and Langlais, 2010). In the task of bilingual lexicon extraction from comparable corpora, the acquisition of translational pairs is mainly based on the Harris' distributional hypothesis (Harris, 1954) which states that words with similar meaning tend to occur in similar contexts (hypothesis that has been extended to the bilingual scenario).

It is well-known that the efficiency of distributional methods heavily depends on the quality and the size of comparable corpora (Morin et al., 2007). If the quality of bilingual lexicons can always be improved by using more data, this is true only if the training data is reasonably well-matched to the desired output (Morin and Hazem, 2014). In the case of specialised comparable corpora, the amount of data is limited and often small (Rapp, 1995; Morin et al., 2007). This presents a problem for distributional methods that fail to extract infrequent single-word terms (SWTs) as well as multi-word terms (MWTs). In the case of MWTs, it has become a standard practice to apply the principle of compositionality on its parts to extract their corresponding translations (Robitaille et al., 2006; Daille and Morin, 2008; Delpech et al., 2012b). While in the case of SWTs, distributional methods often treat them as single tokens without considering their compositional property (Rapp, 1995; Chiao and Zweigenbaum, 2002; Laroche and Langlais, 2010). One can note that a reasonable amount of SWTs are compound terms consisting of a combination of two (or more) lexical elements to form a unit of meaning (Robitaille et al., 2006; Daille and Morin, 2008; Delpech et al., 2012b). To handle derivational morphology, Guevara (2010) and Lazaridou et al. (2013) identify the stem of rare derived words and use its derivational vector to derive the distributional meaning of morphologically complex words from their parts. Delpech et al. (2012a) extract translations of morphologically constructed terms by exploiting a manually constructed translation list of equivalence at the morpheme-level.

In this paper, we apply the compositional property at the morpheme level to automatically build a bilingual list of morphemes (roots and affixes), resource that is not always available and difficult to construct manually. We evaluate our automatic bilingual morpheme extraction based on distributional semantics on two specialized comparable corpora that is the breast cancer and the wind energy corpora for French and English. We propose several ways to model morpheme contexts. We address in the same way neoclassical compounds, quasi-compounds, and prefixed words such as *paramedical*, *immunodeficiency*, and *disappearance*. In this particular case, we first manually split the compound term, then adapt the distributional method by extracting for each part of the compound its corresponding translation, and finally by recomposing the two extracted parts to be the target translation.

To our knowledge, this is the first work on automatic bilingual morphemes extraction from comparable corpora. We show promising results using distributional methods at the root and affix levels and hope that this work can serve as a cornerstone for future work on this task involving more languages. We also confirm that the adapted approach significantly improves bilingual terminology extraction from comparable corpora compared to the baseline system.

2. Related work

Distributional semantic models (DSM) have been successfully used in many natural language processing tasks (Guevara, 2010). Bilingual terminology extraction from comparable corpora for instance, is usually based on the bilingual distributional semantic models (BDSMs) when dealing with single word terms (SWT's) (Rapp, 1999; Gamallo, 2007; Laroche and Langlais, 2010; Morin and Hazem, 2014). To extract a SWT's translation, a similarity measure is applied between the translated context vector of the source SWT and the context vectors of all the target SWTs. The candidates are ranked according to their similarity scores. One of the main problems that encounter distributional methods such as BDSMs is data sparseness. Taking into account the derivational morphology property of a SWT should resolve the latter problem.

Compositional methods which have originally been developed for phrases have been successfully applied by Delpech et al. (2012a) to translate morphologically constructed terms by exploiting a manually constructed translation list of equivalence at the morpheme-level. Guevara (2010) and Lazaridou et al. (2013) have shown that compositional methods improve the quality of monolingual neighbor acquisition. Starting from the assumption that exploiting morphology could improve the quality of distributional semantic models (DSMs) in general and compositional DSMs (cDSM) in particular, and based on the observation that DSMs ignore derivational morphology altogether, Lazaridou et al. (2013) adapted compositional methods to the task of deriving the distributional meaning of morphologically complex words from their parts. They explored the application of compositional distributional semantic models (cDSM) to derivational morphology by adapting several composition methods including the multiplicative model (*mult*) that given input vectors u and v, returns a composed vector c with: $c_i = u_i v_i$; the weighted additive model (*wadd*) where the composed vector c is a weighted sum of the two input vectors: $c = \alpha u + \beta v$ (α and β being two scalars); the full additive model (*fulladd*) where the two vectors u and v are first multiplied by weight matrices and then added as follows: $c = Au + Bv$; the dilation model where one of the input vectors (u or v) is first decomposed into a vector parallel to the other and an orthogonal vector. Before recombining, the parallel vector is dilated by a factor λ giving the following result: $c = (\lambda - 1)\langle u, v\rangle + \langle u, v\rangle v$. In addition, Lazaridou et al. (2013) applyed the lexical function model (*lexfunc*) (Baroni and Zamparelli, 2010) where the distributional representation of one element in a composition is not a vector but a function. They also used the DSM at the stem level as a baseline. Our approach is inspired by the work of Lazaridou et al. (2013) and uses the additive model (*wadd*) to combine several distributional modellings at the morpheme level.

3. Various forms of compounds

Compounding has different forms. First of all, we can talk about "closed compounds" (Macherey et al., 2011) written as single words (e.g, *toolbar*) in contrast to "open compounds", which are space-separated but form a unit of meaning (e.g., *operating system*). We only deal with closed compounds (including hyphen-separated). The major kinds of compounds are native and neoclassical compounds. The first kind includes only native elements, which means not borrowed from another language, suh as *parrot* + *fish* = *parrotfish*. The second kind, neoclassical compounding, combines some elements of Greek or Latin etymological origin, such as *hydro* + *logy* = *hydrology*. Neoclassical elements are not considered as lexical units because they never independently occur in the texts, that is they are always seen in the combined form with other elements (e.g., *biology*) (Amiot and Dal, 2008; Namer, 2009). Each language may assimilate its borrowed neoclassical elements phonologically (but not totally) (Lüdeling, 2006). In other words, a Greek or Latin word undergoes a minimal adaptation before being adopted by a host language. For example, both Fr: *pathie* and En: *pathy* were borrowed from the Greek word *pathos*.

Prefixed words cannot be called compounds in the strict sense of the term because prefixes are not independent lexical units. However some prefixes are very close to the neoclassical roots, compare prefix *bi-* with neoclassical root *uni-* according to (Béchade, 1992). The difference is in their origin (neoclassical roots come from Latin or Greek content words, whereas prefixes come from function words) and the period when they entered into usage (prefixes entered earlier). For our work, we focus on neoclassical compounds and prefixed words, and compounds such as hidden compounds which are at the border between native and neoclassical compounds. Thus, the morphemes under consideration for the distributional analysis are neoclassical elements and prefixes, including elements that are not purely neoclassical elements but look like them, such as the element *radio* in *radiology*.

4. Bilingual morpheme extraction

Our aim is to extract for each source morpheme (root or affix) its corresponding translation in a target language. To do so, we adapted the well-known distributional method to the morpheme level as follows:

1. Each single term of the source and the target language is split into roots or affixes and lexemes. However, many splitting tools are available, either designed for one language such as DeriF (Namer 2003) for the French language, or language independent such as Koehn and Knight (2003) algorithm or COMPOST (Loginova Clouet and Daille, 2014). The single term *abnormal* for instance is split into the prefix *ab-* and the lexeme *normal*;

2. Each lexeme is added to the context vector of its corresponding affix or root according to the co-occurrence of the lexeme with the affix or the root. The lexeme *normal* for instance will be added to the context vector of the prefix *ab-* with the co-occurrence value of *ab-* with *normal*. This corresponds to the occurrence of *abnormal* in the source corpus. This approach is noted $lexem$ for distributional approach using lexemes. At this step four variants of the construction of context vectors can be proposed:

   - In the first variant, we can add the single term (lemma) to the context vector of the affix or the root. Hence, according to the previous example, the single term *abnormal* will be added to the context vector of the prefix *ab-*. This approach is noted lem for distributional approach using lemmas.

- In the second variant and in addition to the lexeme, we can add the single term (lemma). Hence, according to the previous example, both the lexeme *normal* and the single term *abnormal* will be added to the context vector of the prefix *ab-*. This approach is noted $lexem + lem$ for distributional approach using lexemes and lemmas.
- In the third variant and in addition to the lexeme and to the single term, we can add the context words of the lexeme that have been already observed in the corpus. Hence, all the words that appear in the context of *normal* will be added to the context vector of the prefix *ab-*. This approach is noted $Vect(lexem)$ for distributional approach using lexemes and lemmas and the context words of the lexemes.
- In the fourth variant and in addition to the lexeme and to the single term, we can add the context words of the single term that have been already observed in the corpus. Hence, all the words that appear in the context of *abnormal* will be added to the context vector of the prefix *ab-*. This approach is noted $Vect(lem)$ for distributional approach using lexemes and lemmas and the context words of the lemmas.

3. Each lemma or lexeme of the context vector is weighted according to a given association measure such as the point-wise mutual information (Fano, 1961), the discounted odds ratio (Evert, 2005) or the log-likelihood (Dunning, 1993);
4. Each source context vector is translated into the target language using a bilingual dictionary;
5. A similarity measure such as the Cosine (Salton and Lesk, 1968) or the weighted Jaccard (Grefenstette, 1994) is applied between each translated source context vector and all the target vectors;
6. The translation candidates are ranked according to their similarity scores.

The correct translation of the English prefix *ab-* is the French prefix *a-*. Knowing that (automatically thanks to our approach), we can derive from the single word *abnormal* that its French translation is *anormal*. This can be effective if we split *abnormal*, translate each of its parts and then recompose them to obtain the translation.

5. Experiments and Results

We conduct two sets of experiments. The first one aims at evaluating the bilingual morphemes (roots and affixes) extraction for the breast cancer and the wind energy corpora. The second one aims at evaluating the impact of the first experiment on bilingual terminology extraction using the distributional approach.

5.1. Experimental setup

In our experiments we used two French-English comparable corpora. The breast cancer corpus of 500k words and the wind energy corpus of 400k words[1]. The corpora have been normalized through the following linguistic preprocessing steps: tokenization, part-of-speech tagging, and lemmatization. To build our reference lists, we selected French-English morpheme pairs from an existing list of 892 entries [2]. After projection on the bilingual corpus, we obtained 176 morpheme translations from the breast cancer corpus, and 80 morpheme translations from the wind energy corpus. To evaluate the impact of bilingual morpheme extraction on the bilingual terminology extraction task, we built an additional list of morphologically derived terms on the breast cancer corpus. We obtained 32 translations. As bilingual dictionary, we used the French/English ELRA-M0033 resource available from the ELRA catalogue[3]. This resource is a general language dictionary which contains only a few terms related to specialised domains.

Using the distributional method, we chose the log-likelihood (Dunning, 1993) as association measure and weighted Jaccard index (Grefenstette, 1994) as similarity measure. To build the context vectors we chose a 7-window size. Other combinations of parameters were assessed, but the previous parameters turned out to give the best performance.

To evaluate the quality of the system, we used the precision at P1, P5 and P10, we also used the accuracy (Acc.) and the mean average precision *MAP* (Manning et al., 2008).

$$MAP = \frac{1}{|W|} \sum_{i=1}^{|W|} \frac{1}{Rank_i} \tag{1}$$

where $|W|$ corresponds to the size of the evaluation list, and $Rank_i$ corresponds to the ranking of a correct translation candidate i.

5.2. Bilingual morpheme extraction

We applied our adapted distributional approach to each morpheme and evaluate the bilingual morpheme extraction in the two directions that is: from English to French (noted en-fr) and from French to English (noted fr-en). The results are presented in Tables 1 and 2 for breast cancer domain, and in Tables 3 and 4 for wind energy domain.

Table 1 shows the results of English morphemes translation for the breast cancer corpus. We can see that the lemmas-based approach (lem) gives the best results in terms of P1 (30.9%) while the combination of lexemes and lemmas ($lexem + lem$) obtains the best results in terms of P5 (40.3%), P10 (43.2%). Adding the lexemes context information ($Vect(lexem)$) turned out to give the best results in terms of accuracy (61.9%) and adding the lemmas context information obtains the best MAP score with 34.1%.

[1] The Breast cancer and the wind energy corpora can be downloaded from the website http://www.lina.univ-nantes.fr/?Ressources-linguistiques-du-projet.html

[2] The list can be downloaded here http://www.lina.univ-nantes.fr/?Ressources-linguistiques-du-projet,1675.html

[3] http://www.elra.info/

	P1	P5	P10	Acc.	MAP
$lexem$	16.9	35.0	38.0	39.7	23.9
lem	**30.9**	32.7	32.7	32.7	31.8
$lexem + lem$	27.4	**40.3**	**43.2**	43.2	32.7
$Vect(lexem)$	21.0	33.3	39.1	**61.9**	27.8
$Vect(lem)$	28.6	38.5	42.1	60.8	**34.1**

Table 1: Results (%) of morphemes translation for the breast cancer corpus (en-fr)

	P1	P5	P10	Acc.	MAP
$lexem$	19.2	36.2	38.5	41.5	26.2
lem	30.9	32.7	32.7	32.7	31.8
$lexem + lem$	**33.9**	**43.8**	**45.0**	45.0	**38.5**
$Vect(lexem)$	25.1	35.0	38.0	**68.4**	29.8
$Vect(lem)$	29.8	40.3	44.4	67.2	35.2

Table 2: Results (%) of morphemes translation for the breast cancer corpus (fr-en)

Table 2 shows the results of French morphemes translation for the breast cancer corpus. Here, we can see that the combination of lexemes and lemmas ($lexem + lem$) obtains the best results in terms of P1 (33.9%), P5 (43.8%), P10 (45.0%) and MAP (38.5%), while adding the context information of lexemes ($Vect(lexem)$) turned out to give the best results in terms of accuracy with a MAP score of 68.4%, closely followed by the approach based on the lemmas context vectors ($Vect(lem)$) with a MAP score of 67.2%.

	P1	P5	P10	Acc.	MAP
$lexem$	23.7	36.2	38.7	38.7	29.1
lem	**38.7**	40.0	40.0	40.0	39.2
$lexem + lem$	36.2	**43.7**	45.0	45.0	**39.4**
$Vect(lexem)$	25.0	38.7	43.7	63.7	31.0
$Vect(lem)$	31.2	**43.7**	**47.5**	**65.0**	37.2

Table 3: Results (%) of morphemes translation for the wind energy corpus (en-fr)

Table 3 shows the results of English morphemes translation for the wind energy corpus. Similarly to previous results (Tables 1 and 2), we can see that the distributional method based on lemmas (lem) obtains the best precision at P1 (38.7%) while at P5, both $lexem + lem$ and its context information-based method ($Vect(lem)$) obtain the best results (43.7%). The best MAP score is obtained by $lexem + lem$ (39.4%). Finally, $Vect(lem)$ obtains the best accuracy (65%) and P10 (47.5%).

Table 4 shows the results of French morphemes translation for the wind energy corpus. We can see that the lexeme context-based approach ($Vect(lexem)$) obtains the best accuracy (67.5%) while the best P1 (40%) and MAP (40%) results are obtained by the lem approach. Finally, the best P5 (43.7%) and P10 (45%) results are obtained by $lexem + lem$ approach.

	P1	P5	P10	Acc.	MAP
$lexem$	28.7	36.2	37.5	37.5	31.3
lem	**40.0**	40.0	40.0	40.0	**40.0**
$lexem + lem$	36.2	**43.7**	**45.0**	45.0	39.1
$Vect(lexem)$	22.5	37.5	42.5	**67.5**	29.5
$Vect(lem)$	25.0	38.5	43.7	66.2	31.6

Table 4: Results (%) of morphemes translation for the wind energy corpus (fr-en)

We examine more closely the results obtained on the wind energy corpus for the English to French direction. There are elements of which the good translation appears at the first position of the element list whatever are the method and the kind of element, such as the neoclassical element *hypo* translated by *hypo*, or the prefix *un* translated by *ir* or *in*. But generally, the ranks of the element differ according to the context modelling. For example, the translation of the prefix *pre* which is *pré* or *pro* is found at the first position for lem and $lexem + lem$, at the second position for $Vect(lem)$, at the 4^{th} position for $lexem$ and at the 32^{th} position for $Vect(lexem)$. More generally, we can deduce some trends of behaviour according to the context modelling:

- lem ranks the right translation at the first position for 74%, but lem has a medium accuracy (40%);
- $Vect(lem)$ and $Vect(lexem)$ propose the right translation at the first position for 50% for $Vect(lem)$ and 42% for $Vect(lexem)$ with high accuracies (65% and 63.7%);
- $lexem + lem$ offers the best compromise by ranking at the first position 58% of the right translations with a medium accuracy of 45%.

Some elements that are not found by lem modelling such as *sub* with three valid translations *sub*, *sous* or *hypo* are proposed by $Vect(lexem)$ at the first position, by $lexeme$ at the second position, and by the 3 other modellings at the third position. A few elements such as the French suffix *re* appear as translation candidates of almost all English elements. Methods to remove such elements from the list of translation candidates will be useful to improve the ranking, such as the method proposed by Ferret (2013) to remove bad neighbours in distributional analysis.

Overall, we can say that adding information to the basic distributional method applied to lexemes improves bilingual morphemes alignment using comparable corpora. If the $lexem + lem$ method has shown the best results in general for the breast cancer corpus, adding context information has shown better results on the wind energy corpus. These promising results encourage more investigations in the way to exploit the combination of context information.

5.3. Bilingual terminology extraction

In this experiment, we compare the standard distributional approach noted $DistApp$ to the compositional approach based on the proposed bilingual morpheme extraction methods ($lexem$, lem, $lexem + lem$, $Vect(lexem)$ and $Vect(lem)$) for the task of bilingual terminology extraction from comparable corpora. The results are presented in Table 5.

	P1	P5	P10	Acc.	MAP
$DistApp$	6.25	12.5	18.7	**50.0**	10.7
$lexem$	21.8	21.8	21.8	21.8	21.9
lem	**34.3**	**37.5**	**37.5**	37.5	**35.9**
$lexem + lem$	18.7	25.0	25.0	25.0	20.4
$Vect(lexem)$	18.7	25.0	25.0	25.0	20.4
$Vect(lem)$	18.7	25.0	25.0	25.0	20.4

Table 5: Results (%) of bilingual term extraction for the breast cancer corpus (en-fr)

According to Table 5, we can see that using the results of morpheme extraction methods, the compositional approach outperforms the basic standard approach for all the configurations (except in terms of accuracy where the $DistApp$ approach obtains the highest score of 50%). The best results are obtained using the lem approach with a MAP score of 35.9% while the standard approach reaches only 10.7% of MAP score. It is to note that the $lexem$ approach slightly outperform $lexem$+lem, $Vect(lexem)$ and $Vect(lem)$ which obtain the same results. Overall, taking advantage of the morphological information using automatically built lists of morphemes translations is effective to improve bilingual terminology extraction from comparable corpora.

6. Conclusion

We have presented a new method based on distributional semantics to automatically build bilingual translations at the morpheme-level. To our knowledge, this work is the first evaluation of such a task. We hope that our approach can serve as a cornerstone for future works. If additional experiments for other languages than English and French are certainly needed, in the light of this first encouraging results, we can at least conclude that morphological analysis associated to distributional semantics is appropriate for bilingual morphemes alignment as well as for bilingual terminology extraction from comparable corpora. Our experiments have been conducted with a reference segmentation, that is a manual segmentation. We foresee in our next experiments to use as input segmentations provided by a splitting tool. We need to investigate how distributional analysis at the morpheme level deals with erroneous splitting. Loginova-Clouet et al. (2015) compared manual and automatic segmentations for a translation task using a compositional translation. They showed that the results are similar when a precision-oriented segmentation was chosen for the automatic splitting. We hope to reach the same conclusion with automatically built bilingual morpheme translations.

7. References

Amiot, D. and Dal, G. (2008). La composition néoclassique en français et l'ordre des constituants. *La composition dans les langues, Artois Presses Université*, pages 89–113.

Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1183–1193.

Béchade, H.-D. (1992). *Phonétique et morphologie du français moderne et contemporain.* Presses Universitaires de France.

Chiao, Y.-C. and Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *COLING*.

Clouet, E., Harastani, R., Daille, B., and Morin, E. (2015). Compositional translation of single-word complex terms using multilingual splittin. *Terminology. Special Issue: Terminology across languages and domains*, 21(2).

Daille, B. and Morin, E. (2008). An effective compositional model for lexical alignment. In *3rd International Joint Conference on Natural Language Processing (IJ-CLNP)*, Hyderabad, India.

Delpech, E., Daille, B., Morin, E., and Lemaire, C. (2012a). Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 745–762.

Delpech, E., Daille, B., Morin, E., and Lemaire, C. (2012b). Identification of fertile translations in medical comparable corpora: a morpho-compositional approach. *CoRR*, abs/1209.2400.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Evert, S. (2005). *The statistics of word cooccurrences : word pairs and collocations*. Ph.D. thesis, University of Stuttgart.

Fano, R. M. (1961). *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, MA, USA.

Ferret, O. (2013). Identifying bad semantic neighbors for improving distributional thesauri. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 561–571. Association for Computational Linguistics.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In Yarovsky, D. and Church, K., editors, *Proceedings of the 3rd Workshop on Very Large Corpora (VLC'95)*, pages 173–183, Somerset, NJ, US.

Gamallo, P. (2007). Learning bilingual lexicons from comparable english and spanish corpora. In *Proceedings* *of the 11th Conference on Machine Translation Summit (MT Summit XI)*, pages 191–198, Copenhagen, Denmark.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher, Boston, MA, USA.

Guevara, E. (2010). A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37. Association for Computational Linguistics.

Harris, Z. S. (1954). Distributional structure. *Word*.

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of EAC 2003*, Budapest, Hungary.

Laroche, A. and Langlais, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 617–625, Beijing, China.

Lazaridou, A., Marelli, M., Zamparelli, R., and Baroni, M. (2013). Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1517–1526.

Loginova Clouet, E. and Daille, B. (2014). Splitting of Compound Terms in non-Prototypical Compounding Languages. In *Workshop on Computational Approaches to Compound Analysis, COLING 2014*, pages 11 – 19, Dublin, Ireland, August.

Lüdeling, A. (2006). Neoclassical word-formation. In *Keith Brown (ed) Encyclopedia of Language and Linguistics, 2nd Edition*, Oxford, Elsevier.

Macherey, K., Dai, A., Talbot, D., Popat, A., and Och, F. (2011). Language-independent compound splitting with morphological operations. In *Proceedings of ACL 2011*, pages 1395–1404, Portland, Oregon.

Manning, D. C., Raghavan, P., and Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Morin, E. and Hazem, A. (2014). Looking at unbalanced specialized comparable corpora for bilingual lexicon extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1284–1293.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual terminology mining - using brain, not brawn comparable corpora. In *ACL*.

Namer, F. (2009). *Morphologie, lexique et traitement automatique des langues*. Lavoisier, Paris.

Rapp, R. (1995). Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 320–322, Boston, MA, USA.

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 (ACL'99)*, pages 519–526, College Park, MD, USA.

Robitaille, X., Sasaki, Y., Tonoike, M., Sato, S., and Utsuro, T. (2006). Compiling french-japanese terminologies from the web. In *EACL*. The Association for Computer Linguistics.

Salton, G. and Lesk, M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, 15(1):8–36.