Parallel corpora for WordNet construction: machine translation vs. automatic sense tagging

Antoni Oliver and Salvador Climent

Universitat Oberta de Catalunya
Barcelona (Spain)
aoliverg,scliment@uoc.edu
www.uoc.edu

Abstract. In this paper we present a methodology for WordNet construction based on the exploitation of parallel corpora with semantic annotation of the English source text. We are using this methodology for the enlargement of the Spanish and Catalan versions of WordNet 3.0, but the methodology can also be used for other languages. As big parallel corpora with semantic annotation are not usually available, we explore two strategies to overcome this problem: to use monolingual sense tagged corpora and machine translation, on the one hand; and to use parallel corpora and automatic sense tagging on the source text, on the other. With these resources, the problem of acquiring a WordNet from parallel corpora can be seen as a word alignment task. Fortunately, this task is well known, and some aligning algorithms are freely available.

Keywords: lexical resources, wordnet, parallel corpora, machine translation, automatic sense tagging

1 Introduction

WordNet [7] is a lexical database that has become a standard resource in Natural Language Processing research and applications. In WordNet nouns, verbs, adjectives and adverbs are organised in sets of synonyms, the so called synsets. These synsets are connected to other synsets by semantic relations (hiponymy, antonomy, meronomy, troponomy, etc.). For instance, in WordNet 3.0, the synset identified by the offset and pos 06171040-n has two *variants*: *linguistics* and *philology*. Each synset has a gloss or definition, for the synset of the example being: *the humanistic study of language and literature*. It also has a hypernym 06153846-n (*humanistic_discipline, humanities*); and two hyponyms 06171265-n (*dialectology*) and 06178812-n (*lexicology*).

The English WordNet (PWN - *Princeton WordNet*) is being updated regularly, so that its number of synsets increases with every new version. The current version of PWN is 3.1, but we are using in our experiments the 3.0 version.

WordNet versions in other languages are also availabe: in the EuroWordNet project [26] WordNet versions in Dutch, Italian and Spanish have been developed; the Balkanet project [24] developed WordNets for Bulgarian, Greek,

Romanian, Serbian and Turkish; and RusNet [2] for Russian, among others. On the Global WordNet Association[1] website a comprehensive list of WordNets available for different languages can be found.

According to [26], we can distinguish two general methodologies for WordNet construction: (i) the *merge model*, in which a new ontology is constructed for the target language and relations between PWN and this local WordNet are generated; and (ii) the *expand model*, in which English variants associated with PWN synsets are translated following several strategies. In this work and for our purposes we are following this second strategy.

The PWN is a free resource available at the University of Princeton website[2]. Many of the available WordNets for languages other than English are subject to proprietary licenses, although some others are available under free license, for example: Catalan [3], Danish [17], French WOLF WordNet [19], Hindi [21], Japanese [10], Russian [2] or Tamil [18] WordNets among others. The goal of this project is to enlarge and improve the Spanish and Catalan versions of WordNet 3.0 and distribute them under free license.

2 Use of parallel corpora for the construction of WordNets

There are several works using parallel corpora for tasks related to WordNet or WordNet-alike ontologies. In [11], an approach for acquiring a set of synsets from parallel corpora is outlined. Such synsets are derived by comparing aligned words in parallel corpora in several languages. If a given word in a given language is translated by more than one word in several other languages, this probably means that the given word has more than one sense. This assumption also works the other way around. If two words in a given language are translated by only one word in several other languages, this probably means that the two words share the same meaning. A similar idea along with a practical implementation is found in [9], and their results show that senses derived by this approach are at least as reliable as those made by human annotators.

In [8], the Slovene WordNet is constructed using a multilingual corpus, a word alignment algorithm and existing WordNets for some other languages. With the aligned multilingual dictionary, all synsets of the available WordNets are assigned. Of course, some of the words in some of the languages are polysemic, so that more than one synset is assigned. In some of these cases, a word can be monosemic at least in one language, with a unique synset assigned. This synset is used to disambiguate and assign a unique synset in all languages, including Slovene. A very similar methodology is used for French in [19], along with other methods based on bilingual resources.

The construction of an Arabic WordNet using an English-Arabic parallel corpus and the PWN is depicted in [6]. In this parallel corpus the English content words were annotated with PWN synsets.

[1] http://www.globalwordnet.org
[2] http://wordnet.princeton.edu

3 Use of machine translation for the construction of WordNets

Two projects related to WordNet using machine translation systems can be mentioned: the construction of the Macedonian WordNet and the Babelnet project.

In the construction of the Macedonian version of WordNet [20], the monosemic entries are directly assigned using a bilingual English-Macedonian dictionary. For polysemic entries the task can be seen as a Word Sense Disambiguation problem, and thus be solved using a big monolingual corpus and the definitions from a dictionary. However, none of these resources was available. To get Macedonian definitions, PWN glosses were automatically translated into Macedonian using Google Translate. Instead of using a corpus, they took the web as a corpus through the *Google Similarity Distance* [5].

The Babelnet project [15] aims to create a big semantic network by linking the lexicographic knowledge from WordNet to the encyclopedic knowledge of Wikipedia. This is done by assigning WordNet synsets to Wikipedia entries, and making these relations multilingual through the interlingual relations in Wikipedia. For those languages lacking the corresponding Wikipedia entry, the authors propose the use of Google Translate to translate a set of English sentences containing the synset in the Semcor corpus and in sentences from Wikipedia containing a link to the English Wikipedia version. After that, the most frequent translation is detected and included as a variant for the synset in the given language.

In [22], some preliminary experiments on WordNet construction from English machine translated sense tagged corpora are presented. In this paper, this task is presented as a word alignment problem, and some very basic algorithms are evaluated. In [23], these basic algorithms are compared with the Berkeley Aligner for the same task. These papers show that the methodology proposed is promising to build WordNets from scratch, as well as to enlarge and improve existing WordNets.

4 Our approach

4.1 Goal

In this paper we present two approaches for the construction of WordNets based on sense tagged parallel corpora from English to the target language (in our case Spanish). The English part of the corpus must be annotated with PWN synsets. The target part of the corpus does not need to be annotated. To our knowledge, there is no such a corpus freely available for the languages of interest. There are some English sense tagged corpora available, as well as some English and Spanish parallel corpora.

With the available resources, we get a parallel corpora with the English part tagged with PWN synsets in two ways:

– Automatically translating the available English sense tagged corpora into Spanish and Catalan
– Automatically tagging with PWN senses the available English-Spanish parallel corpora

With such a parallel corpus available, the task of constructing a target WordNet can be reduced to a word-alignment task. The relations between the synset in the target WordNet are copied from PWN, assumning that the relations are lingustically and culturally independent from each other.

4.2 Corpora

Sense tagged corpora We have used two freely available sense tagged corpora for English, the tags being the PWN 3.0 synsets:

– The Semcor corpus[3] [14].
– The Princeton WordNet Gloss Corpus (PWGC) [4], consisting of the WordNet 3.0 glosses semantically annotated.

In table 1 we observe the total number of sentences and words in the corpus.

Corpus	Sentences	Words
Semcor	37.176	721.622
PWGC	117.659	1.174.666
Total	154.835	1.896.288

Table 1. Size of the sense tagged corpora

Parallel corpora We have used several subsets of the European Parliament Proceedings Parallel Corpus[5] [12] consisting in the first 200K, 500K amd 1M sentences of the corpus. In table 2 we can observe the number of sentences and words of these subsets and of the full corpus.

4.3 Machine translation

For our experiments we need a machine translation system able to perform good lexical selection, that is, to select the correct target words for the source English sentence. In case of ambiguous words, the system must be able to disambiguate it and choose the correct translation. In our study, other translation errors are less

[3] http://www.cse.unt.edu/∽rada/downloads.html
[4] http://wordnet.princeton.edu/glosstag.shtml
[5] http://www.statmt.org/europarl/

Corpus	Sentences	Words-eng	Words-spa
Full	1.786.594	44.652.439	46.763.624
200K subset	200.000	5.415.925	5.659.496
500K subset	500.000	13.611.548	14.208.128
1M subset	1.000.000	26.830.587	28.121.665

Table 2. Size of the Europarl corpus

important. Therefore we used a statistical machine translation system: Google Translate[6]. In previous works [22] and [23] we also used Microsoft Bing Translator[7] obtaining very similar results.

We did not assess in deep the ability of the system to do a correct lexical selection, but we performed some successful tests. Consider the English word *bank*. According to PWN, it has 10 meanings as a noun, but we will concentrate on only two of them: 09213565n (*sloping land (especially the slope beside a body of water)*) and 08420278n (*a financial institution that accepts deposits and channels the money into lending activities*). The first meaning has three possible variants in Spanish (*margen, orilla, vera*), according to the preliminary version of the Spanish 3.0; whereas the second meaning has only one Spanish variant (*banco*). If we take sentence correspondings to these senses and we translate them with the given MT systems we get:

```
She waits on the bank of the river. Ella espera en la orilla del río.
She puts money into the bank.      Ella pone el dinero en el banco.
```

As we can see, the systems does, at least in certain situations, a good lexical selection. Few references on figures about lexical selection precision for Google Translate can be found in the literature. In [25], a *position-independent word error rate* (PER) of 29.24% is reported for Dutch-English. In [4], a PER of 28.7% is reported for Icelandic-English.

4.4 Automatic sense tagging

For the semantic annotation of the parallel corpora we use Freeling [16]. This linguistic analyser has recently added the UKB algorithm for sense disambiguation, and it is able to tag English texts with PWN 3.0 senses. As we have an English corpus manually tagged with PWN 3.0 senses, we can perform an evaluation of the automatic tagging task. Hence, we have automatically tagged the sense tagged corpus and we have compared each tag with the corresponding one in the manually tagged version of the corpus. In this experiment we got an overall precision of 73.7%.

[6] http://translate.google.com
[7] http://www.microsofttranslator.com/

4.5 Word alignment algorithms

Once we have a parallel corpus sense tagged English - Target Language, the task of deriving the local WordNet can be viewed as a word alignment problem. We need an algorithm capable to select from the following corpus...

```
English:
Then he noticed that the dry wood of the wheels had swollen.
Sense Tagged English:
00117620r he 02154508v that the 02551380a 15098161n of the 04574999n had 00256507v .
Spanish Translation:
Entonces se dio cuenta de que la madera seca de las ruedas se había hinchado.
```

...the following set of relations:

```
00117620r - entonces   02154508v - darse cuenta
02551380a - seco       15098161n - madera
```

Fortunately, word alignment is a well-known task and there are several algorithms available to solve it. In this project we use the Berkeley Aligner[8] [13]. This freely available algorithm performs the alignment task and gives a probability score for each word alignment.

At this stage, we work with the Berkeley Aligner assuming two restrictions: (i) we only detect as a variant for a given synset simple lexical units, that is, no multiwords; and (ii) we only detect one variant for each synset. In a future work we will try to overcome such restrictions. We are using the Berkeley aligner with a combination of MODEL 1 and HMM models with 5 iterations for each model.

5 Evaluation

In this section we present the results of the evaluation of our experiments. Firstly, we present the results for the experiments using machine translation of sense disambiguated corpora. Secondly, we present the evaluation for the experiments using automatic sense tagging of parallel corpora. At the end of this section we present a comparison of the results obtained by each of the two methods.

The evaluation has been carried out automatically using the preliminary version of the Spansih 3.0 WordNet. This evaluation method has a major drawback: since the WordNet of reference is not complete, some correct proposals can be evaluated as incorrect.

The evaluation is performed in an accumulative way, starting with the most frequent synset in the corpus. Results are presented in graphics where the y values represent the accumulate precision and the x values represent the number of extracted synsets.

5.1 Machine translation of sense tagged corpora

In figure 1 we observe the results of the machine translated sense tagged corpus, as well as the evaluation for all alignments and the evaluation for the subsets of alignments with a probability of at least 0.9.

[8] http://code.google.com/p/berkeleyaligner/

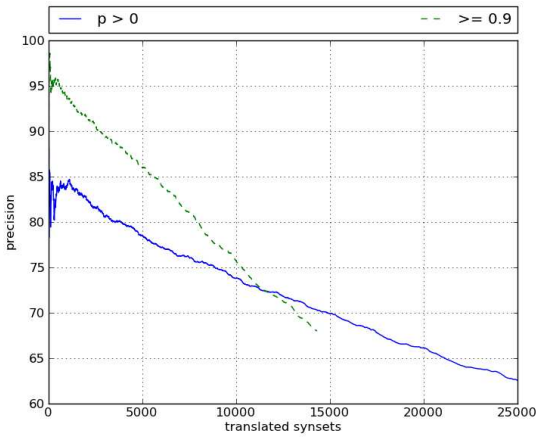

Fig. 1. Precision Berkeley Aligner for the machine translated sense tagged corpus.

With this setting we obtain a variant for 3.880 synsets with a precision of at least 80% and 8.866 with 75%. If we take only the alignments with a probability of at least 0.9 these figures improve, and we obtain one variant for 7.996 synsets with 80% of precision and 10.306 with a precision of 75%.

5.2 Automatic tagging of parallel corpora

In figure 2 we observe the results for the 200K subset of sentences of the Europarl corpus with automatic sense tagging of the English part. Please, note the change of scale when comparing it with figure 1.

In this experiment we obtain poorer results in comparison with results presented in 5.1. If we take into account all the alignments we can not obtain any variant with a precision higher than 75% (in fact, we do not obtain any precision higher than 70%). If we concentrate on alignments with a precision of at least 0.9, we obtain 1.360 variants with a precision of 80% and 1.622 with a precision of at least 75%. These results, compared with results presented in 5.1, suggest that sense tagging is a "more error prone task" than lexical selection in statistical machine translation systems.

Now we are interested in the effects that a bigger corpus could have. In figure 3 we present the results corresponding to aligments with a probability of at least 0.9 for the 200K, 500K and 1M subsets of sentences of the Europarl corpus.

Increasing the size of the corpus has a positive effect in the results. For instance, with the 500K subset of sentences we get variants for 2.355 synsets with a precision of at least 80%, instead of 1.360 corresponding to the 200K subset of sentences. This figure rises up to 3.390 for the 1M subset of sentences.

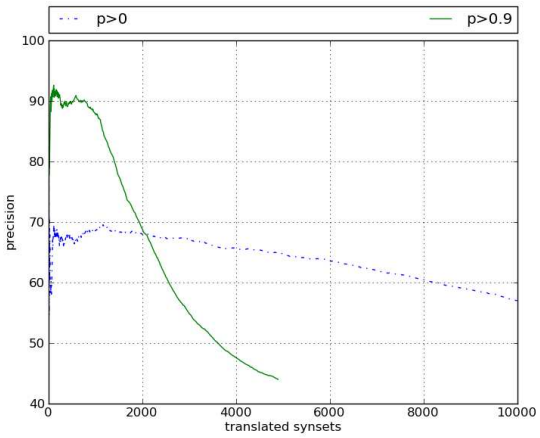

Fig. 2. Precision Berkeley Aligner for the automatically sense tagged 200K sentences Europarl corpus subset.

5.3 Comparison of results for both methods

In figure 4 we observe the results for both corpora: the machine translated manual sense tagged corpus and the automatic sense tagged parallel corpus (1M subset of sentences). As we see, we get better results using the method based on machine translation of sense disambiguated corpora. This suggests that lexical selection errors made by machine translation systems are less important than semantic tagging errors. But we need to further analyse the results in order to find other possible causes.

Another reason may be the different distribution of frequencies in both corpora, as shown in figure 5. As we observe, the frequency of synsets decreases more rapidly in the automatically sense tagged corpus (please, note the log y axis). This can be an additional reason, along with the sense tagging precision (about 73%).

6 Conclusions and future work

In this paper we present a methodology for WordNet construction and enlargement following the expand model based on the exploitation of sense tagged parallel corpora, taking English as a source text. Only the source text needs to be tagged with PWN synsets. With this resource, the task of constructing or enlarging a WordNet can be seen as a word alignment problem. Fortunately, this task is well known and several free algorithms are available. Unfortunately, the required corpus is not easily available. For this reason, we present two proposals for constructing such a corpus in an automatic way: (i) machine translation of a manually sense tagged corpus, and (ii) automatic sense tagging of a manually translated parallel corpus.

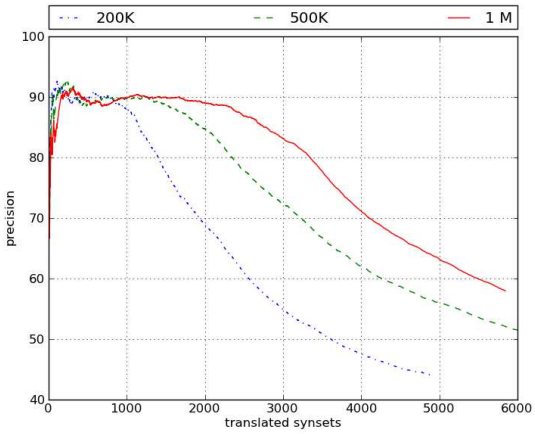

Fig. 3. Comparison of results for differents subsets of the Europarl corpus for $p >= 0.9$

The methodology based on machine translation of sense disambiguated corpora achieves values of precision and number of synsets comparable to methodologies based on bilingual dictionaries English-Spanish [1]. To perform a good comparison we need to further analyse our results, to group variants according to the degree of polysemy. For Spanish, our best algorithm (Berkeley Aligner for $p \geqslant 0.9$) performs better than all the criteria presented in [1] except monosemic-1 criterion. Nevertheless, our proposal performs worse than their combination of criteria, as they obtain 7.131 with a precision higher than 85%, whereas we only obtain 5.562 variants in the same conditions.

The methodology based on automatic tagging of parallel corpora performs much worse. Some reasons can be depicted, but it must be further studied, namely the precision of the sense tagging algorithm and the distribution of synset frequency in the corpus (maybe due to tagging errors or by the corpus typology).

Both methods are prone to errors but our experiments show that the methodology based on automatic sense tagging performs worse. In addition, increasing the size of the corpus has a beneficial effect on the automatic sense tagging method. As it is much easier to construct big parallel corpora than manually tagging monolingual corpora, the increase of the size of the corpus, as well as the selection of more general corpora are aspects to explore in the future.

An important aspect in these experiments that also must be further studied is the selection of the candidates' order. The task is aimed to get the maximum number of variants with the highest possible precision. In the experiments presented in this paper we get the variants in decreasing order of synset frequency in the corpus. Synsets with higher frequency are expected to get the corresponding translated variant with higher probability to be correct, but this is not always the case. In further experiments we plan to take advantage of the information

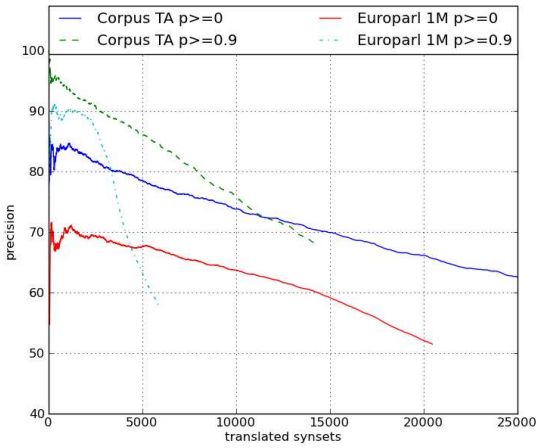

Fig. 4. Comparison of the results for both methods.

given by the alignment algorithm to calculate a function that will allow us to select the variants in a better order.

One drawback of the method based on parallel corpora is the relatively low precision of the automatic sense tagging. To improve the precision we plan to use a multilingual parallel corpora to reduce the degree of ambiguity as depicted in [9]

All the experiments presented in this paper try to get a complete WordNet for a given language. Preliminary local WordNet versions are used only to automatically evaluate the results. In the future, we plan to take advantage of the acquired knowledge to use the preliminary versions to semantically tag the Spanish and Catalan part of the corpus. By doing this we will reduce the difficulty of the task, as some word alignment will be directly done by aligning the same synset ids in both languages.

We also plan to overcome some of the restrictions of the methods presented here: (i) to get more than one variant for each synset, observing the assigned probability of each alignment and taking more than one candidate if probability scores are similar enough; and (ii) to be able to get a lexical unit formed by more of one word as a variant.

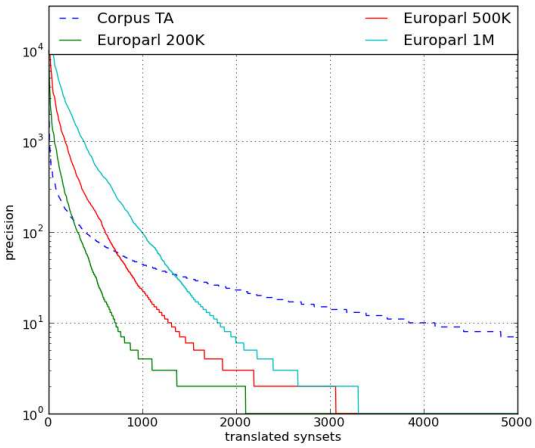

Fig. 5. Comparison of frequency distribution of synsets in both corpora.

References

1. Atserias, J., Climent, S., Farreres, X., Rigau, G., Rodriguez, H.: Combining multiple methods for the automatic construction of multi-lingual WordNets. In: Recent Advances in Natural Language Processing II. Selected papers from RANLP. vol. 97, p. 327–338 (1997)
2. Azarova, I., Mitrofanova, O., Sinopalnikova, A., Yavorskaya, M., Oparin, I.: Russnet: Building a lexical database for the Russian language. In: Workshop on WordNet Structures and Standarisation, and how these affect WordNet Application and Evaluation. pp. 60–64. Las Palmas de Gran Canaria (Spain) (2002)
3. Benítez, L., Cervell, S., Escudero, G., López, M., Rigau, G., Taulé, M.: Methods and tools for building the catalan WordNet. In: In Proceedings of the ELRA Workshop on Language Resources for European Minority Languages (1998), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.63.9020
4. Brandt, M., Loftsson, H., Sigur\thórsson, H., Tyers, F.: Apertium-IceNLP: a rule-based icelandic to english machine translation system. Unpublished paper. Reykjavík: Reykjavik University (2011)
5. Cilibrasi, R.L., Vitanyi, P.M.: The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19(3), 370–383 (2007)
6. Diab, M.: The feasibility of bootstrapping an arabic WordNet leveraging parallel corpora and an english WordNet. In: Proceedings of the Arabic Language Technologies and Resources, NEMLAR, Cairo (2004)
7. Fellbaum, C.: WordNet: An electronic lexical database. The MIT press (1998)
8. Fišer, D.: Leveraging parallel corpora and existing wordnets for automatic construction of the slovene wordnet. In: Proceedings of the 3rd Language and Technology Conference. vol. 7, p. 3–5 (2007)
9. Ide, N., Erjavec, T., Tufis, D.: Sense discrimination with parallel corpora. In: Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8. p. 61–66 (2002)

10. Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., Kanzaki, K.: Development of the japanese WordNet. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008) (2010)
11. Kazakov, D., Shahid, A.: Unsupervised construction of a multilingual WordNet from parallel corpora. In: Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning. p. 9–12 (2009)
12. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT summit. vol. 5 (2005)
13. Liang, P., Taskar, B., Klein, D.: Alignment by agreement. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. p. 104–111. HLT-NAACL '06, Association for Computational Linguistics, Stroudsburg, PA, USA (2006), http://dx.doi.org/10.3115/1220835.1220849, ACM ID: 1220849
14. Miller, G.A., Leacock, C., Tengi, R., Bunker, R.T.: A semantic concordance. In: Proceedings of the workshop on Human Language Technology. p. 303–308. HLT '93, Association for Computational Linguistics, Stroudsburg, PA, USA (1993), http://dx.doi.org/10.3115/1075671.1075742, ACM ID: 1075742
15. Navigli, R., Ponzetto, S.P.: BabelNet: building a very large multilingual semantic network. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. p. 216–225. ACL '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://portal.acm.org/citation.cfm?id=1858681.1858704, ACM ID: 1858704
16. Padró, L., Reese, S., Agirre, E., Soroa, A.: Semantic services in freeling 2.1: Wordnet and UKB. In: Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010) (2010)
17. Pedersen, B., Nimb, S., Asmussen, J., SU00F8rensen, N., Trap-Jensen, L., Lorentzen, H.: DanNet: the challenge of compiling a wordnet for danish by reusing a monolingual dictionary. Language resources and evaluation 43(3), 269–299 (2009)
18. Rajendran, S., Arulmozi, S., Shanmugam, B., Baskaran, S., Thiagarajan, S.: Tamil WordNet. In: Proceedings of the First International Global WordNet Conference. Mysore. vol. 152, p. 271–274 (2002)
19. Sagot, B., Fišer, D.: Building a free french wordnet from multilingual resources. In: Proceedings of OntoLex 2008. Marrackech (Morocco) (2008)
20. Saveski, M., Trajkovski, I.: Automatic construction of wordnets by using machine translation and language modeling. In: 13th Multiconference Information Society. Ljubljana, Slovenia (2010)
21. Sinha, M., Reddy, M., Bhattacharyya, P.: An approach towards construction and application of multilingual indo-wordnet. In: 3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea (2006)
22. Surname1, N., Surname2, N.: The title of the firts autocited paper. In: Proceedings of the First Autocited Conference (1234)
23. Surname1, N., Surname2, N.: The title of the second autocited paper. In: Proceedings of the First Autocited Conference (1234)
24. Tufis, D., Cristea, D., Stamou, S.: BalkaNet: aims, methods, results and perspectives: a general overview. Science and Technology 7(1-2), 9–43 (2004)
25. Vandeghinste, V., Martens, S.: PaCo-MT-D4. 2. report on lexical selection. Tech. rep., Centre for Computational Linguistics - KULeuven (2010), http://www.ccl.kuleuven.be/Projects/PACO/d42.pdf
26. Vossen, P.: Introduction to Eurowordnet. Computers and the Humanities 32(2), 73–89 (1998)