Building a free French wordnet from multilingual resources

Benoît Sagot[1], Darja Fišer[2]
[1]Alpage, INRIA / Université Paris 7
30 rue du Château des rentiers, 75013 Paris, France
[2]Faculty of Arts, University of Ljubljana
Aškerčeva 2, 1000 Ljubljana, Slovenia
E-mail: benoit.sagot@inria.fr, darja.fiser@guest.arnes.si

Abstract

This paper describes automatic construction a freely-available wordnet for French (WOLF) based on Princeton WordNet (PWN) by using various multilingual resources. Polysemous words were dealt with an approach in which a parallel corpus for five languages was word-aligned and the extracted multilingual lexicon was disambiguated with the existing wordnets for these languages. On the other hand, a bilingual approach sufficed to acquire equivalents for monosemous words. Bilingual lexicons were extracted from Wikipedia and thesauri. The results obtained from each resource were merged and ranked according to the number of resources yielding the same literal. Automatic evaluation of the merged wordnet was performed with the French WordNet (FREWN). Manual evaluation was also carried out on a sample of the generated synsets. Precision shows that the presented approach has proved to be very promising and applications to use the created wordnet are already intended.

1. Introduction

The first wordnet was developed for English at Princeton University (PWN). Over time it has become one of the most valuable resources in applications for natural language understanding and interpretation, such as word-sense disambiguation, information extraction, machine translation, document classification and text summarisation and, last but not least, Semantic Web applications (Fellbaum 1998). This initiated the development of wordnets for many other languages apart from English (Vossen 1999, Tufis 2000), which was an important milestone because it enabled the developed resources to be exploited in a multilingual setting as well. Currently, wordnets for more than 50 languages are registered with the Global WordNet Association[1].

While it is true that manual construction of each wordnet produces the best results as far as linguistic soundness and accuracy are concerned, such an endeavour is too time-consuming and expensive to be feasible for most languages. This is why semi- or fully automatic approaches have been proposed. By taking advantage of the existing resources they facilitate faster and easier development of a wordnet.

Apart from the knowledge acquisition bottleneck, another major problem in the wordnet community is the availability of the developed wordnets. Currently, only a handful of them are freely available (Arabic, Hebrew, Irish and Princeton). Although a wordnet for French, the French WordNet (FREWN), has been created within the EuroWordNet project (Vossen 1999), the resource has not been widely used mainly due to licensing issues. In addition, there has been no follow-up work to further extend and improve the core FREWN since the project has ended (Jacquin et al. 2007).

This is why the goal of our experiments presented in this paper was to leverage freely available multilingual resources to automatically construct a broad-coverage open-source wordnet for French called WOLF (Wordnet Libre du Francais)[2].

The rest of the paper is organized as follows: a brief overview of the related work is given in the next section. Section 3 describes the methodology for our experiment. Sections 4 and 5 present and evaluate the results obtained in the experiment and the final section gives conclusions and work to be done in the future.

2. Related work

Automatic techniques for wordnet development can be divided in two approaches: the *merge approach* and the *extend approach* (Vossen 1999). Contrary to the merge approach, according to which an independent wordnet for a certain language is first created based on monolingual resources and then mapped to other wordnets, we have opted for the latter. This model takes a fixed set of synsets from Princeton WordNet (PWN) and translates them into the target language, preserving the structure of the original wordnet. It must be noted here that the extend model presupposes that concepts and semantic relations between them are language independent, at least to a large extent.

Apart from faster and cheaper construction of the lexical resource, the biggest advantage of this approach is that the resulting wordnet is automatically aligned to all other wordnets built on the same principle (e.g. wordnets for Swedish and Russian) and therefore available for use in multi-lingual applications, such as machine translation and cross-language information retrieval.

The cost of the expand model is that the target wordnets are biased by PWN and may, in an extreme case, become completely arbitrary (see Orav & Vider 2004 and Wong 2004).

For example, synset ENG20-09740423-n of PWN contains literals *performer* and *performing artist*. However, there is no word or phrase in French that denotes the concept describing actors, singers and other entertainers collectively. Such cases have been dealt with by providing the closest possible match for the synset and aligning the two wordnets with a near_synonym relation. In this way, the overall structure of straightforward cases remained intact and the exceptions appropriately encoded.

[1] http://www.globalwordnet.org [15.03.2008]
[2] http://wolf.gforge.inria.fr [15.03.2008]

Despite these difficulties, the approach is still attractive due to its much greater simplicity which outweighs the language difference issues This is why the expand model has been adopted in a number of projects, such as the BalkaNet (Tufis 2000) and MultiWordNet (Pianta 2002). It was also used in EWN, including for the construction of FREWN, in which a set of English synsets was automatically translated with a proprietary multilingual semantic database and later manually validated.

Research teams developing wordnets in this setting took advantage of the resources at their disposal, including machine-readable bilingual and monolingual dictionaries, taxonomies, ontologies and others (see Farreres et al. 1998). For the construction of WOLF we have leveraged three different publicly available types of resources: the JRC-Acquis parallel corpus[3], Wikipedia (and other Wikipedia-related resources)[4] and the EUROVOC thesaurus[5].

Equivalents for words that only have one sense in PWN and therefore do not require sense disambiguation were extracted from Wikipedia and the thesaurus in a way, similar to Declerck et al. (2006) and Casado et al. (2005). Roughly 82% of literals found in PWN are monosemous, which means that the bilingual approach suffices for an accurate translation. However, most of these are rather specific and do not belong to the core vocabulary[6].

The parallel corpus was used to obtain semantically relevant information from translations so as to be able to handle polysemous literals as well. The idea that semantic insights can be derived from the translation relation has already been explored by Resnik & Yarowsky (1997), Ide et al. (2002) and Diab (2004). Word-aligned parallel corpora have been used to find synonyms by van der Plas and Tiedemann (2006) and Dyvik (2002). The approach has also yielded promising results in an earlier experiment to obtain synsets for Slovene wordnet (Fišer 2007).

3. Approach

3.1 Alignment approach

In this approach we used used the SEE-ERA.NET corpus (project ICT 10503 RP), a 1.5-million-word subcorpus of JRC-Acquis (Steinberger et al. 2006) in eight languages. Apart from French, we used English, Romanian, Czech and Bulgarian. We used different tools to POS-tag and lemmatize the corpus before word-aligning it with Uplug (Tiedemann 2003). Because word-alignment was done only on single words, the approach was not able to generate any translation equivalents for multi-word expressions.

The output of the word alignment process is a file with word links between word occurrences, associated with the two related word occurrence ids and information on word link certainty.

This allowed us to build bilingual lexicons that include all translation variants of words as well as frequency, POS and word-ids information for each entry. The bilingual lexicons range from 43,024 entries for the Cz-En lexicon to 50,289 for the Cz-Bg one. These bilingual lexicons are then combined into five multilingual lexicons. They contain between 49,356 (Fr-Ro-Cz-Bg-En) to 59,019 entries (Fr-Cz-Bg-En). A few entries from the Fr-Cz-Bg-En lexicon are shown in Table 1. Obviously, not all these entries are correct; errors may appear for several reasons, such as tagging, lemmatization, or alignment problems. However, most of these errors are eliminated by the next stage of the process.

frq	**pos**	**Fr**	**Cs**	**Bg**	**En**
18	n	droit	právo	законодателство	law
56	n	droit	právo	право	law
4	n	loi	právo	закон	law
4	n	loi	právo	законодателство	law
6	n	loi	právo	право	law
33	n	loi	zákon	закон	law
8	n	loi	zákon	закона*	law
19	n	législation	právo	законодателство	law
7	n	législation	právo	право	law
4	n	législation	předpis	законодателство	law

Table 1: Translation variants of the English literal *law* from the Fr-Ro-Cs-Bg-En lexicon[7].

At the next stage the goal was to assign a synset id to each lexicon entry. To achieve this, we gathered the set of all possible synset ids assigned to each lexicon entry in all languages (apart from the French one, of course) by comparing it with the corresponding BalkaNet wordnet (Tufis 2000). This is possible because all BalkaNet wordnets use the same synset ids as PWN 2.0. We could then compute the intersection of ids for all languages. The result contains all synset ids that are shared among all non-French lexicon entries. We then assigned these synset ids to their French equivalent. Let us illustrate this by taking the French word *droit*, which is polysemous in French (possible English translation equivalents are: *right*, *law*, *droit*, *royalty*, *entitlement*, *claim*). As shown by Table 1, 56 of its occurrences were aligned with *právo* in Czech, *право* in Bulgarian and *law* in English. The intersection of all sets of synset ids containing the word in wordnets for each individual language contains only the synset id ENG20-05791721-n. It is therefore assigned to those occurrences of the French word *droit* (see Table 2). It is one of the correct synsets for this word (defined in PWN as *the branch of philosophy concerned with the law and the principles that lead courts to make the decisions they do*).

[3] http://langtech.jrc.it/JRC-Acquis.html [15.03.2008]

[4] http://www.wikipedia.org [15.03.2008]

[5] http://europa.eu/eurovoc [15.03.2008]

[6] When we refer to the core vocabulary in this paper, we have in mind all literals corresponding to concepts that are included in the BalkaNet Basic Concept Sets (Tufis 2000). There are three categories of basic synsets, BCS1 being the most fundamental one.

[7] 4-uples occurring 3 times or less are not shown. The literal marked by an asterisk comes from lemmatization errors.

Multiple languages disambiguate polysemous lexicon entries and eliminate most alignment errors. It is rather unlikely that the same polysemy occurs in many different languages or that alignment errors lead to a non-empty intersection. Therefore, the intersection of all possible senses in each language is likely to output only the correct synset.

Fr: ***droit***	**Cs:** ***právo***	**Bg:** ***право***	**En:** ***law***
droit	ENG20-06129345-n	ENG20-04893549-n	ENG20-00577416-n
	ENG20-05559593-n	ENG20-04888072-n	ENG20-05529208-n
	ENG20-05791721-n	ENG20-07928837-n	ENG20-05531141-n
	ENG20-04617988-n	ENG20-00577416-n	**ENG20-05791721-n**
	ENG20-07928837-n	**ENG20-05791721-n**	ENG20-06129345-n
		ENG20-01000872-n	ENG20-07712371-n
		ENG20-04881053-n	ENG20-07928837-n
		ENG20-04617988-n	

Table 2: Word sense disambiguation and sense assignment for French lexicon entries

Applied to the above-mentioned multilingual lexicons, this technique yielded five different sets of synsets with at least one French literal. They include between 1,338 (Fr-Ro-Cs-Bg-En) and 5,073 (Fr-Ro-En) synsets. Because the preprocessing stages, such as tagging, lemmatization and word-alignment were not perfect, it is expected that the synsets created in this way will inherit some of the errors, of course. However, the approach covers polysemous literals from the core vocabulary, which the translation approach, described in the next section, cannot handle.

3.2 Translation approach

We used the following freely available bilingual resources to translate monosemous literals from the PWN 2.0 into French:

- Wikipedia[8] is an on-line multilingual collaborative encyclopaedia. We used it to build a bilingual Fr-En lexicon (314,713 entries) by following to inter-wiki links that relate two articles on the same topic in French and English. We improved and extended this lexicon with a quick analysis of article bodies (capitalization, synonyms extraction, preliminary extraction of definitions).
- The French Wiktionary and its English counterpart[9] are lexical companions to Wikipedia that contain definitions of words as well as some additional information, including their translations into other languages. We used them to create a bilingual lexicon with 24,464 (from the English Wiktionary) and 24,873 entries (from the French Wiktionary).
- Wikispecies[10] is a taxonomy of living species which include both Latin standard names and (for common species) vernacular terms. This allowed us to identify 129,509 language-independent Latin terms as well as French equivalents for 2,648 of these Latin terms.
- Eurovoc[11] is a multilingual thesaurus that is used for classification of EU documents. Version 4.2 of the thesaurus is a structured list of 6,802 descriptors and their equivalents in 21 languages, including many multi-word expressions.

All the bilingual lexicons we extracted from these resources were used to translate monosemous PWN literals. We obtained sets of synsets of different sizes: 18,273 from Wikipedia, 6,848 from Wikispecies, 6,215 and 4,363 from the French and English Wiktionary, and 1,319 from Eurovoc. Translations of the monosemous literals are very accurate and include many multi-word expressions, which was a serious limitation of the alignment approach. Also, they mostly contain specific, non-core vocabulary.

3.3 Merging the results

In the end, synsets obtained from both approaches were merged. If the same synset was created from more than one resource (e.g. from a multilingual lexicon that was extracted from the word-aligned corpus and from a bilingual lexicon that was extracted from Wikipedia), all their unique literals were retained along with the information on the source of the generated synset. This enabled us to perform a simple heuristic filtering according to the reliability of each source, on the diversity of sources that assign a given literal to a given synset and on frequency information (for sources from the alignment approach).

Automatic induction of synsets inevitably leads to gaps in the hierarchy. Because we are aware of the importance of the conceptual density and hierarchy preservation principles for applications (Tufis 2000), we inherited the structure and relations of the missing synsets from PWN 2.0. Empty synsets will need to be addressed in the future. But for the time being, in case an application runs into an empty synset, it can still use the relation information to access a more general or more specific concept. Other language-independent information (e.g. POS, domain, semantic relations) was inherited from PWN.

4. Results

WOLF currently contains 32,351 non-empty synsets that include 38,001 unique literals (see Table 3). This is substantially more than the number of synsets present in FREWN (22,857 in the original resource, but 22,121 once FREWN synsets are mapped to PWN 2.0 synsets). This is directly related to the high number of monosemous PWN literals in non-core synsets (119,528 out of 145,627), that the translation approach was able to handle well.

WOLF contains all four parts of speech that are normally coded in wordnets, while there are only nouns and verbs in FREWN. The most common literals in WOLF are nouns (34,827 vs. 14,618 in FREWN). They are followed by adjectives (1,521 vs. 0 in FRWEN), verbs (979 vs. 3,777 FREWN), and adverbs (664 vs. 0 in FREWN).

[8] http://www.wikipedia.org [15.03.2008]

[9] http://www.wiktionary.org [15.03.2008]

[10] http://species.wikimedia.org [15.03.2008]

[11] http://europa.eu/eurovoc [15.03.2008]

	PWN 2.0	WOLF	WOLF/PWN	FREWN	FREWN/PWN
All synsets	115,424	32,351	**28.0%**	22,121	**19.2%**
BCS1	1,218	870	71.4%	1,211	99.4%
BCS2	3,471	1,668	48.0%	3,022	87.1%
BCS3	3,827	1,801	47.1%	2,304	60.2%
non-BCS	106,908	28,012	26.2%	15,584	14.6%
nominal	79,689	25,559	35.8%	17,381	21.8%
verbal	13,508	1,544	11.5%	4,740	35.1%
adjectival	18,563	1,562	8.4%	0	0.0%
adverbial	3,664	676	18.4%	0	0.0%

Table 3: Quantitative data about WOLF in comparison to PWN and FRWN.

Average polysemy in WOLF is 1.21 synsets per literal (10.5% of literals are polysemous, including 1.2% of multiword literals). In PWN 2.0, average polysemy stands at 1.74 synsets per literal, and 1.39 in FREWN. Coverage of the core vocabulary in WOLF was checked on Base Concept Sets and then compared to FREWN. As Table 3 shows, the core vocabulary in FREWN is denser that in WOLF but the latter has a reasonable coverage of BCS senses as well (71.4% of BCS1, 51.0% of all BCS). It also shows, unsurprisingly, that the more basic the synset, the more likely it is to have been built with the alignment approach.

5. Evaluation

The quality of the resource we created was evaluated automatically as well as manually. In automatic evaluation we compared the resulting wordnet to FREWN and computed f-measure. For a better insight into the problems of our techniques we took a closer look at a representative sample of literals that were not assigned a 100% precision in automatic evaluation. The errors we identified in manual evaluation were classified into several categories.

3.1 Automatic evaluation

FREWN was used as a gold standard to compute precision and recall of sense assignment in WOLF. The most straightforward approach for evaluation of the quality of the obtained wordnet would be to compare the generated synsets with the corresponding synsets from FREWN. But in this way we would be penalizing the automatically induced wordnet for missing literals, which are not part of the vocabulary of the corpus or the bilingual resources that were used to generate the synsets. Instead we opted for a somewhat different approach by comparing literals in the gold standard and in the automatically induced wordnet with regard to which synsets they appear in. This information was used to calculate precision, and recall. Precision gives the number of synset ids assigned to a literal by both wordnets according to the number of synset ids assigned by WOLF. Recall gives the number of synset ids assigned to a literal by both wordnets according to the number of synset ids assigned by FREWN. Results are shown in Table 4.

It must be noted here, however, that literals translated with Wikipedia have a 93,0% precision compared to FREWN. Since the majority of non-BCS synsets are populated from Wikipedia, most synsets that go beyond the coverage of FREWN are of very high quality. Moreover, if a literal appears in a particular synset in WOLF whereas it does not in FREWN, this does not necessarily mean that there is an error in WOLF but it is also possible that FREWN may be incomplete. We therefore selected a sample of 100 literals that were not assigned a 100% precision in automatic evaluation and looked at them by hand as described below.

POS	WOLF/align		WOLF/transl		WOLF/total	
	Prec	Rec	Prec	Rec	Prec	Rec
n	77.2%	68.7%	82.6%	74.9%	**80.4%**	**74.5%**
v	65.8%	54.7%	54.8%	35.8%	**63.2%**	**52.5%**
n+v	74.6%	65.4%	78.8%	69.6%	**77.1%**	**70.3%**

Table 4: Precision and recall of WOLF compared to FREWN for nominal and verbal synsets[12].

3.1 Manual evaluation

A set of randomly selected 100 literals for which WOLF and FREWN show discrepancies was checked by hand. They correspond to 183 literal-synset pairs. We checked manually whether the generated literal-synset pairs are correct or not. We classified errors into several categories, according to the relationship between the literal and the synset it is associated with:

- it is semantically close to the synset (hypernym, hyponym, near-synonym; e.g. *absence* in the synset {*lack, deficiency, want*}),
- it is semantically related (any other kind of semantic relation; e.g. *abri* in the synset {*penthouse*}),
- it is morphologically related (it is part of a compound which would have been correctly assigned to the synset if word alignment was not restricted to single words, or it is a morphologically different form of an otherwise correct literal; e.g. *affaire* in the synset {*things*}, whereas the plural

[12] FREWN does not contain any adjectives or adverbs which could therefore not be evaluated automatically.

form *affaires* would be correct; *aisance* in the synset {*toilet, lavatory, lav, can, john, privy, bathroom*} whereas the compound *cabinet d'aisances* would have been correct),
- it is not related at all (because of alignment and/or disambiguation error; e.g. *abattre* in the synset {*excavate, dig up, turn up*}).

POS	**n**	**v**	**adj**	**adv**	**all**
in FREWN	76 68%	33 46%	0 0%	0 0%	109 60%
not in FREWN					
correct	16	18	*4*	*0*	38
sem. close	10	6	*0*	*0*	17
sem. related	2	6	*0*	*0*	7
morph. related	2	0	*0*	*0*	2
not related	5	5	*0*	*0*	10
total	**111**	**68**			**183**
total correct (WOLF prec.)	**92** **83%**	**51** **75%**	***4***	***0***	**147** **80%**

Table 5. Manual evaluation of WOLF[13].

The results for different POS are shown in Table 5. Approximately 50% of discrepancies are literals that are missing in FREWN synses rather than errors in WOLF. Unsurprisingly, the least problematic synsets are those lexicalizing specific concepts (such as *hippopotamus*, *kitchen*) and the most difficult ones were those containing highly polysemous words describing vague concepts (e.g. *face* which as a noun has 13 different senses in PWN or *place* which as a noun has 16 senses). For a more detailed evaluation, including the resource-by-resource evaluation and resource confidence ranking, see Fišer and Sagot (submitted).

6. Conclusions and future work

The paper has presented a methodology to combine several freely available resources in order to generate a wordnet for a new language. The evaluation of the results shows that the proposed approach is promising from quantitative as well as qualitative aspects. However, precision of the automatically generated synsets drops as ambiguity of words increases, thus affecting the core vocabulary in the developed resource the most. This means that a systematic manual revision of the automatically generated synsets is necessary in order increase the overall quality of WOLF and turn it into a useful resource for NLP applications. Synsets from Base Concept Sets are already being edited by our students.

In addition to this, we intend to extend automatic techniques in order to improve the coverage of WOLF. In particular, we plan to use word sense disambiguation techniques such as those described in Ruiz (2005) to assign synset ids to polysemous Wikipedia entries.

We also plan to extend the scope of WOLF's use and evaluation. In particular, we want to use it for parsing disambiguation and information retrieval purposes. Not only will this validate the usefulness of the resource, it will also enable a more application-oriented evaluation of its relevance and the necessary refinement.

7. References

Casado, R. M., E. Alfonseca, and P. Castells (2005): Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia. In: *Natural Language Processing and Information System*s: *10th International Conference on Applications of Natural Language to Information Systems*, NLDB 2005, Alicante, Spain, June 15-17, 2005.

Christine Jacquin, Emmanuel Desmontils, Laura Monceaux (2007): French EuroWordNet Lexical Database Improvements. In: *Proceedings of CICLing 2007*, pp. 12—22.

Declerck, Thierry, Asunción Gómez Pérez, Ovidiu Vela, Zeno Gantner, David Manzano-Macho (2006): Multilingual Lexical Semantic Resources for Ontology Translation. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy, 24-26 May 2006.

Diab, Mona (2004): 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.

Dyvik, Helge (2002). *Translations as semantic mirrors: from parallel corpus to wordnet*. Revised version of paper presented at the ICAME 2002 Conference in Gothenburg.

Farreres, Xavier, G. Rigau, H. Rodrguez (1998): Using WordNet for Building WordNets. In: *Proceedings of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, Canada.

Fellbaum, Christiane (1998): *WordNet: An Electronic Lexical Database*. MIT Press.

Fišer, Darja (2007). Leveraging parallel corpora and existing wordnets for automatic construction of the Slovene wordnet. In: *Proceedings of the 3rd Language and Technology Conference*, LTC07, Poznan, Poland, October 3-5 2007.

Fišer, Darja, Benoît Sagot (submitted): *Combining multiple resources to build reliable wordnets*.

Ide, Nancy, Tomaž Erjavec, Dan Tufis (2002): Sense Discrimination with Parallel Corpora. In: *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, pp. 54--60.

Orav, Heili and Kadri Vider (2004): Concerning the Difference Between a Conception and its Application in the Case of the Estonian WordNet. In: *Proceedings of the Second Global WordNet Conference*, pp. 285--290, Brno, Czech Republic, January 20-23, 2004.

[13] Figures in italics have to be considered with caution, given the small amount of corresponding data.

Pianta, Emanuele, L. Bentivogli, C. Girardi: MultiWordNet (2002): developing an aligned multilingual database. In: *Proceedings of the First International Conference on Global WordNet*, Mysore, India, January 21-25, 2002.

Resnik, Philip, David Yarowsky (1997): A perspective on word sense disambiguation methods and their evaluation. In: *ACL-SIGLEX Workshop Tagging Text with Lexical Semantics: Why, What, and How?* April 4-5, 1997, Washington, D.C., pp 79--86.

Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, Dániel Varga (2006): The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy, 24-26 May 2006.

Tiedemann, Jörg (2003): *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*, Doctoral Thesis. Studia Linguistica Upsaliensia 1.

Tufis, Dan (2000): BalkaNet - Design and Development of a Multilingual Balkan WordNet. In: *Romanian Journal of Information Science and Technology Special Issue* (Volume 7, No. 1-2).

van der Plas, Lonneke, Jörg Tiedemann (2006): Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity. In: *Proceedings of ACL/COLING 2006*.

Vossen, Piek (ed.) (1998): *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.

Wong, Shun Ha Sylvia (2004): Fighting Arbitrariness in WordNet-like Lexical Databases - A Natural Language Motivated Remedy. In: *Proceedings of the Second Global WordNet Conference*, pp. 234--241, Brno, Czech Republic, January 20-23, 2004.