Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages

Scott Piao[1], Paul Rayson[1], Dawn Archer[2], Francesca Bianchi[3], Carmen Dayrell[1], Mahmoud El-Haj[1], Ricardo-María Jiménez[4], Dawn Knight[5], Michal Křen[6] , Laura Löfberg[1], Rao Muhammad Adeel Nawab[7], Jawad Shafi[1], Phoey Lee Teh[8], Olga Mudraya[9]

[1]Lancaster University, UK; [2]Manchester Metropolitan University, UK; [3]University of Salento, Italy, [4]Universitat Internacional de Catalunya, Spain; [5]Cardiff University, UK; [6]Charles University in Prague, Czech Republic; [7]COMSATS Institute of Information technology, Pakistan; [8]Sunway University, Malaysia; [9]Independent Researcher, UK

E-mail: s.piao@lancaster.ac.uk, p.rayson@lancaster.ac.uk,d.archer@mmu.ac.uk, francesca.bianchi@unisalento.it, c.dayrell@lancaster.ac.uk, m.el-haj@lancaster.ac.uk, rmjimenez@uic.es, knightd5@cardiff.ac.uk, michal.kren@ff.cuni.cz, l.lofberg@lancaster.ac.uk, adeelnawab@ciitlahore.edu.pk, jawadshafi@ciitlahore.edu.pk, phoeyleet@sunway.edu.my, olga.mudraya@ntlworld.com

Abstract

The last two decades have seen the development of various semantic lexical resources such as WordNet (Miller, 1995) and the USAS semantic lexicon (Rayson et al., 2004), which have played an important role in the areas of natural language processing and corpus-based studies. Recently, increasing efforts have been devoted to extending the semantic frameworks of existing lexical knowledge resources to cover more languages, such as EuroWordNet and Global WordNet. In this paper, we report on the construction of large-scale multilingual semantic lexicons for twelve languages, which employ the unified Lancaster semantic taxonomy and provide a multilingual lexical knowledge base for the automatic UCREL semantic annotation system (USAS). Our work contributes towards the goal of constructing larger-scale and higher-quality multilingual semantic lexical resources and developing corpus annotation tools based on them. Lexical coverage is an important factor concerning the quality of the lexicons and the performance of the corpus annotation tools, and in this experiment we focus on evaluating the lexical coverage achieved by the multilingual lexicons and semantic annotation tools based on them. Our evaluation shows that some semantic lexicons such as those for Finnish and Italian have achieved lexical coverage of over 90% while others need further expansion.

Keywords: semantic lexicon, multilingual lexicons, semantic annotation, lexical coverage, corpus annotation tool, multilingualism

1. Introduction

This paper reports on the construction of large-scale multilingual semantic lexicons for twelve languages, which have been designed to employ a unified semantic classification scheme and to form the knowledge bases for an automatic multilingual semantic annotation system.

Over the past two decades, we have seen the development of various semantic lexical resources such as WordNet[1] and the USAS semantic lexicon (Rayson et al., 2004), which are playing an important role in the areas of natural language processing, corpus-based language studies and information system development. More recently, increasing efforts have been devoted to extending the semantic frameworks of existing lexical knowledge resources to cover more languages: see, e.g., EuroWordNet (Vossen, 1998), and the Global WordNet[2] etc.

Creating multilingual resources around a common unified framework is an important endeavour as it supports not only monolingual research in those languages, but also cross-lingual and multilingual research directly. Such aspects are becoming even more important in a number of application scenarios and research settings such as speech-to-speech machine translation, cross-lingual text reuse and plagiarism detection, summarisation and information retrieval.

Our work contributes towards the goal of constructing larger-scale and higher-quality multilingual semantic lexical resources and corpus annotation tools. In this paper, we focus on the lexical coverage of multilingual semantic lexicon resources developed by the UCREL[3] team based at Lancaster University, in collaboration with partner research teams. This multilingual resource is an extension of the core USAS English semantic lexicon, which provides a knowledge base for the USAS semantic annotation system (Rayson et al., 2004), and has been continuously expanded during a number of projects (Löfberg et al., 2005; Mudraya et al., 2006; Piao et al., 2015).

Different from many existing lexical resources, which are built as independent lexical knowledge bases, our semantic lexicons form components of the USAS system, in which the lexicons and software framework are integrated seamlessly to provide a software system for automatically annotating text. In addition, they all apply a common unified lexicographically inspired framework via the same semantic taxonomy across all languages. This enables us to port the USAS software rapidly by generating the semantic lexicons for new languages with a pre-defined information and presentation format. The lexical items in the resource include single word lexemes and multiword expressions (MWEs), and are classified using part-of-speech information and a semantic field categorisation scheme (Archer et al., 2004), which can support various cross-lingual studies such as multilingual

[1] https://wordnet.princeton.edu/wordnet

[2]For details of Global WordNet, see http://globalwordnet.org/wordnets-in-the-world

[3] http://ucrel.lancaster.ac.uk/

semantic analysis, content analysis, and corpus linguistics in multiple languages. These multilingual semantic lexicons have been used to extend the USAS semantic tagger to other languages. As such, their lexical coverage is an important factor when evaluating the performance of the multilingual semantic annotation of USAS.

2. Construction of Multilingual Semantic Lexicons

The current USAS multilingual semantic lexicons cover twelve languages, namely Finnish, Russian, Italian, Chinese, Portuguese, Spanish, Dutch, Czech, Urdu, Malay, Arabic and Welsh. Primarily they have been created using the following four approaches:

1) Automatically translating the core English semantic lexicon using bilingual dictionaries and other publicly available lexicons;
2) Using crowdsourcing methods to clean and expand the automatically generated lexicons;
3) Where possible, using bilingual parallel corpora to align words across languages, thereby allowing the application of 1) and 2);
4) Using machine translation tool to directly translate existing lexicons into new languages.
5) Manually cleaning and curating the lexicons whenever possible.

For example, the Finnish and Russian lexicons were built mainly manually with the help of corpus analysis tools (Löfberg et al., 2005; Mudraya et al. 2006), while the Italian, Chinese and Portuguese lexicons were built by automatically transferring the semantic tags of the English semantic lexicon entries to their translation equivalents in these languages using bilingual dictionaries and other freely available bilingual lexicons (see Piao et al., 2015 for a full description of this process and resources used).

In the case of Czech, a large parallel corpus InterCorp (Čermák and Rosen, 2012) (version 8) was used to automatically extract a Czech--English bilingual dictionary. In particular, we used its manually sentence-aligned fiction core data, as well as journalistic packages from PressEurop/VoxEurop and Project Syndicate. Next, an automatic word-to-word alignment with GIZA++ (Och and Ney, 2003) was carried out, followed by a summarisation of the resulting word pairs as possible translation equivalents, where the summarisation process kept the POS differences of the words, i.e. if a word has multiple POS tags the word-tag pairs form separate entries in the translation lexicon. Finally, the Czech semantic lexicon was generated automatically by transferring the English semantic tags to Czech via the aligned word translation equivalents.

The multilingual WordNet was also tested to extract a bilingual lexicon for the Malay language by porting the semantic lexicons via synset IDs, where the quality of the translation remains an issue. Wherever possible, the different methods were combined. For example, for Spanish, we used a combination of automatically generated and manually compiled sub-lexicons. Mohamed et al. (2013) describe the creation process for the Arabic version.

An important factor concerning the scale and quality of our multilingual semantic lexicon construction is the availability of bilingual/multilingual lexical and corpus resources. Because we rely on open and freely accessible resources, we are faced with the issue of data scarcity for some languages. For instance, there are very few open-access Welsh/English bilingual lexical resources available. The only one we have currently found suitable for our purpose is Nodine's (2003) bilingual lexicon, which was built some time ago and may not accurately reflect the latest language usage. Another important issue is to find an effective approach for cleaning and improving the automatically generated semantic lexicons at scale. For this purpose, the crowdsourcing method has been under investigation and will be reported in future papers.

The semantic lexicons of different languages are in various stages of development, as shown in Table 1. Some of the lexicons have undergone extensive manual cleaning and improvement whereas some others are largely generated automatically. For example, the semantic categories of the lexical entries of the entire Finnish and Russian lexicons have been manually checked and corrected; the Italian lexicons have been generated with extensive manual intervention and correction; and the remaining lexicons have been largely generated automatically. As our work progresses, all of the lexicons of these languages will be continuously improved. With regards to our focus in this paper - the lexical coverage of the lexicons - the quality of the semantic classification of lexical entries is obviously important but is a secondary issue here and will be evaluated elsewhere.

Language	Automatic vs manual creation	Cleaning and improvement	Semantic tagger exists?
Finnish, Russian	Manually compiled	Full manual cleaning and improvement	yes
Italian, Spanish	Auto & manually generated	Extensive manual cleaning	yes
Arabic	Auto generated	Extensive manual cleaning	no
Chinese, Czech, Dutch, Portuguese	Auto generated	Minor or no manual correction	yes
Malay, Urdu, Welsh	Auto generated	Initial experimental stage, not lemmatised	no

Table 1: Development stages of semantic lexicons for different languages.

The sizes of the various lexicons for different languages are dependent on the availability of resources. Larger lexicons have already been collected for Malay, Czech and Chinese, for example, but we have only recently commenced the lexicon construction for some other languages. Note, however, that a larger size of lexicon does not necessarily indicate that the lexicon is directly suitable for application in the automatic tagging software as yet, because some of them are not lemmatised or

cleaned, and hence may contain "noisy" lexical items. Table 2 lists the size of the current lexicons for each language. Again, as our work progresses, all of the lexicons of these languages will be continuously expanded.

Language	Single word entries	Multiword expressions
Arabic	31,154	0
Chinese	64,541	19,048
Czech	28,161	0
Dutch	4,220	0
Finnish	46,225	4,422
Italian	13,098	5,622
Malay	64,863	0
Portuguese	13,499	1,781
Russian	17,443	713
Spanish	3,665	0
Urdu	1,765	235
Welsh	2,744	0

Table 2: Semantic lexicon sizes for 12 languages (sorted alphabetically).

It can be observed from the above table that we have initially focussed on extending the coverage of the single word lexicons rather than MWEs. Uncategorised lists of single words are essentially trivial to collect from corpora, and more amenable to porting their candidate semantic categories via bilingual word lists. Creating lists of semantically meaningful multiword expressions for each new language is much more challenging in order to cover the variety of types of MWEs e.g. phrasal verbs, compound nouns and true non-compositional idiomatic expressions. Candidate lists for some MWE types can be bootstrapped using hybrid rule-based and statistical approaches (Piao et al., 2005) but automatic assignment of potential semantic tags to them is more problematic than for single words, especially for idiomatic expressions.

3. Experiment

In order to estimate the lexical coverage of the existing semantic lexicons built for the twelve languages, we carried out a comparative evaluation of the lexical coverage using a large-scale multilingual corpus as the test data, as described in the following sub-sections.

3.1 Test Data

As the test data of our evaluation, we chose the HC multilingual corpus (Christensen, 2014). This choice was based upon several important requirements. The corpus is a large freely available multilingual corpora (Tan et al., 2014), which provides very recent language data and reflects language features in the genres of journalism, blogging and social media (Twitter). In addition, it provides collections of newspapers, blogs and Twitter messages for over 60 languages, covering all the languages included in our evaluation. Considering the "noise" expected from the Twitter messages, we have focussed, in particular, on the newspaper and blogging data for this evaluation. In fact, the blogging data also contain some "noise", particularly code switching (containing phrases and sentences of different languages), which negatively affects the lexical coverage. For each language, we extracted two sets of data from the newspapers and blogging sections of the HC corpus respectively, with each set containing a million words (types containing inflected words). Interestingly, the HC dataset provides the national or regional versions for Spanish and Portuguese, i.e. European vs. non-European versions, and hence it allows us to evaluate coverage separately for these varieties or dialects. A similar issue applies to other languages such as Arabic, which may affect the coverage as well as the accuracy of part-of-speech taggers chosen to embed within the semantic taggers.

3.2 Evaluation results

The aim of our evaluation is to estimate the percentage of the words in the running texts that can be tagged with semantic tags by the USAS taggers (currently available for eight languages) or contained in the semantic lexicons (for Urdu, Malay, Welsh and Arabic). We divided the languages into two groups: those with semantic taggers available versus those with no semantic tagger software developed yet. Firstly, for those languages for which semantic taggers have been built, we tagged the test corpus data with the semantic taggers and calculated percentage of the words that were assigned with meaningful semantic tags. Table 3 below lists the evaluation results for the eight languages, including four language variants of European (EU)/Brazilian (BR) versions of Portuguese and European (EU)/South America (SA) versions of Spanish. As shown, our lexicons and taggers obtained encouraging coverages for several languages, with the top coverage of 95.93% for Finnish.

Language	Blogs (%)	News (%)	Average (%)
Finnish	95.98	95.89	95.93
Italian	91.14	89.34	90.24
Czech	87.95	86.05	86.99
Russian	84.93	86.66	85.79
Chinese	82.98	79.36	81.17
Portuguese (EU)	76.79	77.47	77.13
Portuguese (BR)	76.11	77.75	76.93
Dutch	61.55	59.87	60.71
Spanish (EU)	57.81	55.73	56.77
Spanish (SA)	57.20	56.11	56.65

Table 3: Lexical coverage on the HC test data assessed using semantic taggers (sorted by average coverage).

Next, for the four languages for which semantic tagger software have not been developed yet, the semantic lexicons were used as a dictionary and we estimated the proportion of the words (types) in the test corpus that were contained in the lexicons. As lemmatisers were not used

for these languages, the lexical coverages for them may be skewed and underestimated. Nevertheless, we assume that these figures can provide an approximate estimate of their lexical coverages. Table 4 below shows the results.

Language	Blogs (%)	News (%)	Average (%)
Arabic	86.43	91.33	88.88
Urdu	86.26	84.21	85.24
Malay	53.83	54.91	54.37
Welsh	27.05	24.73	25.89

Table 4: Lexical coverages estimated by directly matching lexicon entries (semantic taggers not available yet).

In order to estimate the reliability of our evaluation results, we further divided the test data for each language/language variant and text type (blogs versus news) into ten sub-divisions. For each of the language and type categories, we tested the lexical coverage on each of the sub-divisions, then calculated the statistical deviation score. Our assumption is that if the lexical coverages of the individual sub-sections close to each other, or have a small statistical variation score, then it would indicate that the semantic taggers and semantic lexicons have stable lexical coverages on similar types of text and vice versa. Table 5 below displays the lexical coverages of the sub-sections and the variation scores for each category, where the codes ***b*** and ***n*** denote the text types of blogs and newspapers respectively. As shown in the table, the lexicons achieved small variation scores, with the maximum of 0.0047 on Urdu news data. Such small deviation scores indicate that our semantic lexicons have rather stable lexical coverage across different sub-sections of the test data.

Lang.		Div1	Div2	Div3	Div4	Div5	Div6	Div7	Div8	Div9	Div10	Deviation
Finnish	b	0.9590	0.9574	0.9607	0.9587	0.9605	0.9598	0.9602	0.9597	0.9611	0.9610	0.0012
	n	0.9597	0.9602	0.9586	0.9585	0.9590	0.9581	0.9600	0.9585	0.9581	0.9576	0.0009
Italian	b	0.9112	0.9122	0.9113	0.9142	0.9088	0.9128	0.9115	0.9116	0.9108	0.9092	0.0015
	n	0.8939	0.8901	0.8927	0.8944	0.8957	0.8947	0.8932	0.8924	0.8946	0.8915	0.0017
Arabic	b	0.8610	0.8650	0.8620	0.8670	0.8620	0.8650	0.8670	0.8630	0.8600	0.8690	0.0030
	n	0.9130	0.9160	0.9110	0.9110	0.9150	0.9140	0.9140	0.9130	0.9130	0.9120	0.0020
Czech	b	0.8775	0.8797	0.8783	0.8801	0.8815	0.8791	0.8806	0.8789	0.8803	0.8784	0.0012
	n	0.8582	0.8593	0.8609	0.8604	0.8649	0.8606	0.8587	0.8600	0.8605	0.8611	0.0018
Russian	b	0.8524	0.8447	0.8512	0.8474	0.8509	0.8504	0.8514	0.8478	0.8451	0.8509	0.0028
	n	0.8643	0.8643	0.8671	0.8640	0.8674	0.8675	0.8673	0.8660	0.8691	0.8682	0.0018
Urdu	b	0.8578	0.8592	0.8572	0.8584	0.8675	0.8699	0.8615	0.8686	0.8626	0.8629	0.0046
	n	0.8445	0.8409	0.8406	0.8323	0.8493	0.8375	0.8435	0.8443	0.8462	0.8419	0.0047
Chinese	b	0.8307	0.8294	0.8297	0.8280	0.8303	0.8294	0.8310	0.8300	0.8297	0.8289	0.0009
	n	0.7923	0.7941	0.7959	0.7968	0.7959	0.7910	0.7895	0.7944	0.7927	0.7926	0.0023
Portuguese (EU)	b	0.7646	0.7683	0.7679	0.7666	0.7685	0.7673	0.7694	0.7652	0.7678	0.7726	0.0022
	n	0.7753	0.7777	0.7769	0.7754	0.7719	0.7769	0.7726	0.7749	0.7712	0.774	0.0022
Portuguese (BR)	b	0.7591	0.7640	0.7651	0.7603	0.7592	0.7588	0.7634	0.7598	0.7595	0.7610	0.0023
	n	0.7798	0.7796	0.7770	0.7726	0.7812	0.7779	0.7752	0.7768	0.7762	0.7783	0.0025
Dutch	b	0.6164	0.6172	0.6126	0.6139	0.6164	0.6169	0.6185	0.6169	0.6127	0.6128	0.0022
	n	0.5951	0.6005	0.5975	0.599	0.5990	0.6004	0.5977	0.5974	0.6005	0.5992	0.0017
Spanish (EU)	b	0.5814	0.5785	0.5777	0.5760	0.5802	0.5775	0.5762	0.5797	0.5777	0.5759	0.0019
	n	0.5587	0.5581	0.5547	0.5584	0.5595	0.5535	0.5577	0.5615	0.5541	0.5570	0.0025
Spanish (SA)	b	0.5721	0.5721	0.5688	0.5724	0.5689	0.5726	0.5731	0.5744	0.5724	0.5728	0.0018
	n	0.5609	0.5608	0.5636	0.5625	0.5554	0.5615	0.5636	0.5610	0.5618	0.5595	0.0023
Malay	b	0.5385	0.5405	0.5388	0.5364	0.5370	0.5387	0.5410	0.5387	0.5363	0.5368	0.0017
	n	0.5470	0.5500	0.5447	0.5502	0.5500	0.5494	0.5511	0.5481	0.5488	0.5507	0.0020
Welsh	b	0.2692	0.2705	0.2712	0.2716	0.2698	0.2680	0.2705	0.2743	0.2708	0.2695	0.0017
	n	0.2456	0.2484	0.2473	0.2482	0.2476	0.2464	0.2463	0.2457	0.2497	0.2474	0.0013

Table 5: Lexical coverage deviation across ten sub-divisions of test data for each language

The lexical coverage of the frequent words can also help us to assess the practical usefulness of the lexicons for general language analysis. To examine this feature of the lexicons, we calculated the coverage of our lexicons for the 500 and 1,000 most frequent words of the entire test data of each language involved in our experiment. As shown in Table 6, our lexicons achieved over 90% of coverage for seven languages, with Arabic obtaining a full coverage. Meanwhile, the results also show that the semantic lexicons of some other languages, including Welsh, Malay, Portuguese and Spanish, need to expand significantly.

To sum up, over a decade of continuous development and expansion, the USAS multilingual semantic lexicons have become a significant lexical knowledge resource that already achieves high lexical coverages for several

languages. Integrated into the USAS semantic annotation software, they will provide useful tools for corpus annotation and NLP research.

Language	Top 500 (%)	Top 1,000 (%)
Arabic	100.0	100.0
Italian	100.0	99.5
Finnish	100.0	99.4
Urdu	100.0	85.3
Russian	97.8	96.5
Chinese	97.8	95.1
Czech	92.2	92.4
Dutch	85.2	83.4
Portuguese (BR)	81.4	81.4
Portuguese (EU)	80.4	80.3
Spanish (EU)	82.8	71.6
Spanish (SA)	82.0	72.3
Malay	62.2	62.5
Welsh	8.8	6.5

Table 6: Lexical coverage on most frequent 500 and 1000 words (sorted by coverage)

4. Conclusion

In this paper, we reported on the lexical coverage of multilingual semantic lexicons that have been built for the USAS semantic annotation system. In our experiment, the semantic lexicons demonstrated encouraging lexical coverage for the majority of the twelve languages in our lexicons. Coupled with semantic annotation software, they can provide useful tools for various cross-lingual and multilingual studies.[4] We will continue to expand and improve the multilingual semantic lexicons to cover more languages and improve their quality, aiming to develop a large scale multilingual corpus semantic analysis system. In future work, we will turn our attention to the generation of candidate MWE lists for multiple languages as well as evaluating crowdsourcing methods and inter-rater agreement for the process of checking lists of potential semantic tags on each word and MWE.

The multilingual lexicons are available to download from https://github.com/UCREL/Multilingual-USAS and are available under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

5. Acknowledgements

This work has been supported by the UCREL research centre, Lancaster University, UK via Wmatrix software licence fees. We also acknowledge the support of the Centre for Corpus Approaches to Social Science, ESRC grant reference ES/K002155/1. The Finnish and Russian semantic lexicon construction were supported by the EU BENIDICT (The New Intelligent Dictionary) Project and the UK EPSRC ASSIST (Automated Semantic Assistance for Translators) Project (ref. EP/C004574) respectively, and the Welsh semantic lexicon construction is supported by the UK ESRC/AHRC CorCenCC (The National Corpus of Contemporary Welsh) Project (ref. ES/M011348/1).

[4] For examples from English, see http://ucrel.lancaster.ac.uk/wmatrix/

6. References

Archer, D., Rayson, P., Piao, S., McEnery, T. (2004). Comparing the UCREL semantic annotation scheme with lexicographical taxonomies. In *Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004)*. Lorient, France. Volume III, pp. 817--827.

Christensen, H. (2014). HC corpora. http://www.corpora.heliohost.org/.

Čermák, F., Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3), 411–427.

Löfberg, L., Piao, S., Nykanen, A., Varantola, K., Rayson, P., and Juntunen, J.P. (2005). A semantic tagger for the Finnish language. In *Proceedings of the Corpus Linguistics Conference 2005*. Birmingham, UK.

Miller, G.A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, Vol. 38, No. 11: 39-41.

Mohamed, G., Hardie, A, and Potts, A. (2013). AraSAS: a semantic tagger for Arabic. In *Proceedings of The Second Workshop on Arabic Corpus Linguistics, Lancaster*. URL: http://eprints.lancs.ac.uk/69928/

Nodine, M.H. (2003). Welsh to English lexicon. https://www.cs.cf.ac.uk/fun/welsh/LexiconWE.html

Mudraya, O.V., Babych, B.V., Piao, S., Rayson, P. and Wilson, A. (2006). Developing a Russian semantic tagger for automatic semantic annotation. In *Proceedings of the International Conference of Corpus Linguistics – 2006*. St.-Petersburg, Russia, pp. 290-297.

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

Piao, S., Bianchi,F., Dayrell, C., D'Egidio, A. and Rayson, P. (2015). Development of the multilingual semantic annotation system. In *Proceedings of The 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015)*. Denver, Colorado, USA.

Piao, S., Rayson, P., Archer, D. and McEnery, T. (2004). Evaluating lexical resources for a semantic tagger. In *Proceedings of The 4thInternational Conference on Language Resources and Evaluation (LREC 2004)*.Lisbon, Portugal,pp. 499-502.

Piao, S., Rayson, P., Archer, D., McEnery, T. (2005) Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech and Language*, 19 (4), pp. 378 – 397.

Rayson, P., Archer, D., Piao, S. and McEnery, T. (2004). The UCREL semantic analysis system. In *Proceedings of LREC-04 Workshop: Beyond Named Entity Recognition Semantic Labeling for NLP Tasks*.Lisbon, Portugal, pp. 7-12.

Tan, L., Zampieri, M., Ljubeši´c, N., Tiedemann, J.

(2014). Merging comparable data sources for the discrimination of similar languages: The DSL corpus collection. In *Proceedings of Building and Using Comparable Corpora Workshopof LREC'14*. Reykjavik, Iceland.

Vossen, P. (ed.) (1998). EuroWordNet: *A multilingual database with lexical semantic networks*.Kluwer Academic Publishers. ISBN 0792352955.