Development of the Multilingual Semantic Annotation System

Scott Piao
Lancaster University
Lancaster
UK
s.piao@lancaster.ac.uk

Francesca Bianchi
University of the Salento
Lecce
Italy
francesca.bianchi@unisalento.it

Carmen Dayrell
Lancaster University
Lancaster
UK
c.dayrell@lancaster.ac.uk

Angela D'Egidio
University of the Salento
Lecce
Italy
angela.degidio@unisalento.it

Paul Rayson
Lancaster University
Lancaster
UK
p.rayson@lancaster.ac.uk

Abstract

This paper reports on our research to generate multilingual semantic lexical resources and develop multilingual semantic annotation software, which assigns each word in running text to a semantic category based on a lexical semantic classification scheme. Such tools have an important role in developing intelligent multilingual NLP, text mining and ICT systems. In this work, we aim to extend an existing English semantic annotation tool to cover a range of languages, namely Italian, Chinese and Brazilian Portuguese, by bootstrapping new semantic lexical resources via automatically translating existing English semantic lexicons into these languages. We used a set of bilingual dictionaries and word lists for this purpose. In our experiment, with minor manual improvement of the automatically generated semantic lexicons, the prototype tools based on the new lexicons achieved an average lexical coverage of 79.86% and an average annotation precision of 71.42% (if only precise annotations are considered) or 84.64% (if partially correct annotations are included) on the three languages. Our experiment demonstrates that it is feasible to rapidly develop prototype semantic annotation tools for new languages by automatically bootstrapping new semantic lexicons based on existing ones.

1 Introduction

In this paper, we report on an experiment to develop prototype semantic annotation tools for Italian, Chinese and Brazilian Portuguese based on an existing English annotation tool. Over the last twenty years, semantic lexical resources and semantic annotation tools, such as EuroWordNet (Vossen, 1998) and USAS (Rayson et al., 2004), have played an important role in developing intelligent NLP and HLT systems. Various applications of semantic annotation systems and annotated corpus resources have been reported, including empirical language studies at the semantic level (Rayson et al. 2004; Ooi et al., 2007; Beigman Klebanov et al., 2008; Potts and Baker, 2013) and studies in information technology (Volk, et al., 2002; Nakano et al, 2005; Doherty et al., 2006; Chitchyan et al., 2006; Taiani et al., 2008; Gacitua et al., 2008) among others.

While various semantic annotation tools are available for monolingual analysis, particularly for English, there are few such systems that can carry out semantic analysis of multiple languages with a unified semantic annotation scheme. We aim to address this issue by extending an existing English semantic annotation tool (Rayson et al., 2004) to cover a range of languages.

The USAS semantic annotation tool mentioned above adopts a lexical semantic classification scheme derived from Tom McArthur's Longman Lexicon of Contemporary English (McArthur, 1981), which consists of 21 main discourse fields and 232 sub-fields, such as “social actions, states and processes” and “emotion” etc. It also uses a set

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1268–1274,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

of auxiliary codes, such as ***m/f*** (male/female), **+/-** (positive/negative) etc. For example, it tags “happy” and “sad” with “E4.1+” and “E4.1-” respectively, indicating positive and negative sentiment. It also identifies many types of multi-word expressions, such as phrasal verbs, noun phrases, named entities and true non-compositional idioms, and annotates them with single semantic tags since this is highly significant for identifying contextual meaning. Recent applications of the USAS tagger include analysis of literary language (Balossi, 2014), the language of psychopaths (Hancock et al, 2013) and scientific deception (Markowitz and Hancock, 2014). There would be obvious benefits if such a semantic tool could cover a wide range of languages. Efforts have been made to port the existing semantic annotation system to other languages (Finnish and Russian) (Löfberg et al., 2005; Mudraya et al., 2006), so a prototype software framework could be used. However, manually developing semantic lexical resources for new languages from scratch is a time consuming task. In this experiment, we examine the feasibility of rapidly bootstrapping semantic lexical resources for new languages by automatically translating existing English semantic lexicons using bilingual dictionaries. We developed prototype semantic annotation tools for Italian, Chinese and Brazilian Portuguese based on automatically generated semantic lexicons. Our evaluation of the tools shows that it is feasible to rapidly develop prototype semantic tools via the aforementioned automatic method, which can be improved and refined manually to achieve a high performance.

2 Related Work

There exist various tools that can semantically annotate multilingual texts, including GATE (Cunningham et al., 2011) and KIM (Popov et al., 2003) which, combined together, provide multilingual semantic annotation functionalities based on ontologies. Freeling (Padró et al., 2012) provides multilingual annotations such as named entity recognition and WordNet sense tagging. Recent developments in this area include Zhang and Rettinger’s work (2014) in which they tested a toolkit for Wikipedia-based annotation (wikification) of multilingual texts. However, in the work described here we employ a lexicographically-informed semantic classification scheme and we perform *all-words* annotation. In terms of porting tools from one language to another by translating lexicons, Brooke et al. (2009) obtained poor results from a small dictionary in cross-linguistic sentiment analysis.

3 Generating Multilingual Semantic Lexicons by Automatic Mapping

The USAS tagger relies heavily on the semantic dictionary as its knowledge source, so the main task in the development of our prototype semantic annotation tools for new languages was to generate semantic lexicons, both for single word and multiword expressions (MWE), in which words and MWEs can be associated with appropriate semantic tags. For this purpose, our approach involves mapping existing English semantic lexicons into target languages in order to transfer the semantic tags across translation equivalents. The entries of the English semantic lexicons are classified under the USAS semantic annotation scheme (Archer et al., 2004), which consists of 21 major semantic categories that are further divided into 232 subcategories.

In order to translate the English semantic lexicons into other languages, we needed a bilingual lexicon for each of the target languages, Italian, Chinese and Portuguese in our particular case. For this purpose, we first used two corpus-based frequency dictionaries compiled for Chinese (Xiao et al., 2009) and Portuguese (Davies and Preto-Bay, 2007), which cover the 5,000 most frequent Chinese and Portuguese words respectively. These dictionaries provided high-quality manually edited word translations. In addition, we used large English-Italian and English-Portuguese bilingual lexicons available from FreeLang site (http://www.freelang.net/dictionary) as well as an English-Chinese bilingual word list available from LDC (Linguistic Data Consortium). Compiled without professional editing, these bilingual word lists contain errors and inaccurate translations, and hence they introduced noise into the mapping process. However, they provided wider lexical coverage of the languages involved and complemented the limited sizes of the high-quality dictionaries used in our experiment. Table 1 lists the bilingual lexical resources employed for translating the English lexicons into each of the three languages involved in our experiment.

Language	Lexical resources
Italian	English-Italian FreeLang wordlist (33,700 entries);
Chinese	Chinese/English dictionary (5,000 entries); LDC Eng-Chi bilingual wordlist (110,800 entries)
Portuguese	Portuguese/English dictionary (5,000 entries); English-Portuguese (Brazilian version) FreeLang wordlist (20,980 entries)

Table 1: Bilingual lexical resources used.

The semantic lexicon translation process mainly involves transferring semantic tags from an English lexeme to its translation equivalent/s. For instance, given a pair of word/MWE translations, one of which is English, if the English headword is found in the English semantic lexicon, its semantic categories are passed to its translation equivalents. For the high-quality formal dictionaries, this approach worked very well in our experiment, thanks to the accurate translations and explicit part-of-speech (POS) information provided by such resources.

With the bilingual word lists from FreeLang and LDC, however, this translation process was not straightforward. Firstly, most of the entries of the word lists do not contain any POS information. To avoid losing any potentially relevant semantic tags, we have to consider all possible POS categories of each English headword, and the same applies to their translation equivalents. For example, the English headword "advance" has four possible *C7* POS tags (*JJ*-adjective, *NN1*-singular noun, *VV0*-base form of verb, *VVI*-infinitive verb) in the English semantic lexicon with different semantic categories including *N4* (linear order), *A9-* (giving), *M1* (moving, coming and going), *A5.1* (evaluation: good/bad), *A2.1* (affect: modify, change), Q2.2 (speech acts), *S8+* (helping), *Q2.1* (speech etc: communicative), although with some overlap, as shown below (in each line, the first code is a POS tag and the following ones denote USAS semantic categories[1]):

advance	*JJ N4*
advance	*NN1 A9- M1 A5.1+/A2.1*
advance	*VV0 M1 A9- Q2.2 A5.1+/A2.1*
advance	*VVI M1 S8+ A9- A5.1+/A2.1 Q2.1*

In such a case, for each of the possible translation equivalents of the word "advance", these four types of POS tags and their corresponding semantic tags need to be assigned to their corresponding translations in the target languages. Obviously this would lead to passing wrong and redundant semantic tags to the translation equivalents. Nevertheless, we have to accept such noise in order to increase the chances of obtaining correct semantic tags, as it would be easier to remove redundant/incorrect semantic tags than searching for missing ones in the manual improvement stage.

Another major challenge in the translation process was the mapping between the POS tagsets employed by different lexical resources and tools. Even for the same language, different lexicons and tools can employ different POS tagsets. For example, different Portuguese POS tagsets are used by the Portuguese frequency dictionary and the POS TreeTagger (Schmid, 1994). To bridge between the different POS tagsets, we designed a simplified common POS tagset for each language, into which other tags can be mapped. For example, the Portuguese POS tagset was simplified into 12 categories "adj, adv, det, noun, pnoun, verb, pron, conj, intj, prep, num, punc". Because a single semantic category tends to span similar POS categories, e.g. present/past/progressive tense of verbs, simplification of POS tagsets generally does not affect semantic annotation accuracy.

After applying all the resources and automatic mapping described above, we obtained approximately 38,720, 83,600 and 15,700 semantic lexicon entries for Italian, Chinese and Portuguese respectively. Our initial evaluation involved direct manual checking of these bootstrapped lexicons. For example, 5,622 Italian MWE entries and 1,763 Italian single word entries have been manually corrected. For the Chinese lexicon, the most frequent words were identified using the Chinese word frequency list of Internet Corpus (Sharoff, 2006), and the semantic tags of about 560 entries related to the most frequent words were manually corrected. For Portuguese, about 900 lexicon entries were manually checked.

The manual improvement mainly involves three processes: a) filtering lexicon entries having wrong POS tags, b) selecting correct semantic tags from candidates, c) adding missing semantic tags. The amount of effort needed depends on the quality of the bilingual dictionaries. For example, from the automatically generated 900 Chinese entries containing the most frequent (also highly ambiguous) words, 505 entries were selected after the POS filtering. In addition, 145 of them were improved by

[1] For definitions of the POS and semantic tags, see websites http://ucrel.lancs.ac.uk/claws7tags.html and http://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf

adding missing semantic tags. Table 2 shows the sizes of the current lexicons.

Language	Single word entries	MWE entries
Italian	33,100	5,622
Chinese	64,413	19,039
Portuguese	13,942	1,799

Table 2: Sizes of current semantic lexicons.

4 Architecture of Annotation System

Based on the multilingual semantic lexicons described in the previous section, prototype semantic taggers were built for the three languages by deploying the lexicons into the existing software architecture, which employs disambiguation methods reported by Rayson et al. (2004). A set of POS tagging tools were incorporated to pre-process texts from the target languages. The TreeTagger (Schmid, 1994) was used for Italian and Portuguese, and the Stanford POS tagger (Toutanova et al., 2003) was used for Chinese. These tools and semantic lexicon look-up components form pipelines to annotate words in running texts. Figure 1 shows the architecture of the software framework.

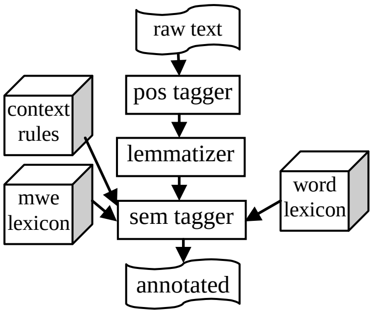

Figure 1: Architecture of the semantic tagger.

5 Evaluation of Prototype System

Following the initial manual evaluation of the prototype semantic taggers described in section 3, we then carried out larger scale automatic evaluations using a set of sample corpora. We conducted two complementary types of evaluations: lexical coverage and annotation precision. The lexical coverage is a particularly interesting metric for our evaluation, as we expect this is where an automatic approach can make significant contribution to the development of annotation systems. On the other hand, high annotation precision normally entails manual improvement of the lexical resources or a period of training on manually tagged corpora.

For the lexical coverage evaluation, three reference corpora were chosen: PAISÀ Italian corpus (Borghetti et al., 2011), LCMC Corpus (Lancaster Corpus of Mandarin Chinese) (McEnery and Xiao, 2004) and Lacio-Ref Portuguese corpus (Aluisio et al., 2003). Because PAISÀ and Lacio-Ref corpora are too large for our purpose, we extracted subsections of about 1.5 million Italian words and 1.7 million Portuguese words from them.

For the evaluation, we annotated the corpus data using the annotation tools of the corresponding target languages, and examined what percentage of the words were assigned with semantic tags. Punctuation marks were excluded in this evaluation process. Table 3 shows the statistics of the evaluation for each language.

Language	Number of words	Tagged words	Lexicon coverage (%)
Italian	1,479,394	1,265,399	85.53
Chinese	975,482	786,663	80.64
Portuguese	1,705,184	1,251,579	73.40
Average			79.86

Table 3: Lexical coverage of the semantic taggers.

As shown in the table, the annotation tools achieved an average lexical coverage of 79.86% over the three languages, with Italian having the highest coverage of 85.53% and Portuguese the lowest coverage of 73.40%. Due to the different types of data in the three sample corpora, this result is not conclusive. Homogeneous corpus data from all of the three languages will be needed to make more reliable comparison of the lexical coverage. Considering that the tools were built based on only three bilingual lexical resources over a short period of time, such lexical coverage is encouraging. This result also demonstrates that, if sufficiently large bilingual lexicons become available; our approach can potentially achieve high lexical coverage.

Next we conducted an evaluation of the precision of the prototype tools. We randomly selected sample texts for each language as follows. Italian sample texts were selected from domains of press, contemporary literature and blogs; Chinese sample texts from press, reviews and fiction; Portuguese sample texts from press and fiction. In the evaluation, we annotated the sample texts using the prototype annotation tools and manually checked the precision among the annotated words. We used two metrics: correctly tagged and partially correctly tagged. With the current tools, a word can be assigned with multiple candidate semantic tags. The first evaluation metric refers to the cases where the first candidate tag is correct, whereas the other metric refers to the cases where the other tags in the list are correct or closely related to the true word sense. Table 4 shows the statistics of the evaluation.

Lan.	Sample text size	Tagged words	Correct	Partially correct
Ita	4,510	3,266	1,826 (55.91%)	672 (20.58%)
Chi	1,053	813	616 (75.76%)	97 (11.93%)
Port	1,231	953	787 (82.58%)	68 (7.14%)
Avg			71.42%	13.22%

Table 4: Evaluation of precision.

As shown in the table, the Portuguese tagger obtained the highest first-tag precision (82.58%), while the Italian tagger produced a precision (55.91%) significantly lower than others. However, if we include the partially correct annotations, the precision scores become more consistent: 76.49%, 87.69% and 89.72% for the three languages respectively, with an average precision of 84.64%. We also estimated recall based on the numbers of tokens of the sample texts and those tagged correctly/partially correctly, obtaining 55.39%, 67.71% and 69.46% for Italian, Chinese and Portuguese respectively. Such a fairly close range of the precision and recall values indicates that our approach to developing prototype semantic annotation tools can be expected to achieve stable results across various languages, although we need larger-scale evaluations to draw a conclusion. It is worth noting that, although the recall is still low, these taggers are starting to approach the precision of the English system at 91% (Rayson et al., 2004).

Our further error analysis revealed that the main causes of the errors include the homonym translations (e.g. *bank* as river bank vs. money bank), translation errors and missing of the translation words in the English semantic lexicons. For example, the Chinese word “爸爸” (father) has a number of synonymous English translation equivalents in the bilingual lexicon: *dad* (with semantic tag *S4m*), *baba, da, dada, daddy* (*S4m*), *father* (*S4m S9/S2m*), *papa* (*S4m*). It is also translated into *presence* (*M6, A3+, S1.1.3+, S1.2, S9*) by mistake. Among the correct English translations, *baba, da, dada* (transliteration) are not included in the English semantic lexicons. Making things worse, *da* is a homonym which is classified as a discourse marker of exclamation (*Z4*) in English lexicons. Our current automatic process collects all the semantic tags derived from the English translation counterparts found in the bilingual lexicon and assigns them to the Chinese word “爸爸”, resulting in an erroneous entry as shown below:

爸爸 *noun M6 A3+ S1.1.3+ S1.2 S9 S4/B1 S4m S9/S2.2m Z4*

In order to resolve such cases, we will need to consider contexts of each translation word pairs’ usage via parallel or comparable corpora.

6 Conclusion and Future Work

In this paper, we have investigated the feasibility of rapidly bootstrapping semantic annotation tools for new target languages[2] by mapping an existing semantic lexicon and software architecture. In particular, we tested the possibility of automatically translating existing English semantic lexicons into other languages, Italian, Chinese and Brazilian Portuguese in this particular case. Our experiment demonstrates that, if appropriate high-quality bilingual lexicons are available, it is feasible to rapidly generating prototype systems with a good lexical coverage with our automatic approach. On the other hand, our experiment also shows that, in order to achieve a high precision, parallel/comparable corpus based disambiguation is needed for identifying precise translation equivalents, and a certain amount of manual cleaning and improvement of the automatically generated semantic lexicons is indispensible. We are continuing to improve the multilingual semantic taggers and extend them to cover more languages, such as Spanish and Dutch, aiming to develop a large-scale multilingual semantic annotation and analysis system. We also intend to perform task-based evaluation of the manually checked versus automatically generated lexicons.

Acknowledgments

We would like to thank Prof. Elena Semino, Prof. Yufang Qian and Dr. Richard Xiao for their contributions to our research. This work is supported by the UCREL research centre and the ESRC Centre for Corpus Approaches to Social Science (CASS), ESRC grant reference ES/K002155/1, both at Lancaster University, UK.

[2] The results are available at http://ucrel.lancs.ac.uk/usas/

References

Aluisio, Sandra M., Gisele Pinheiro, Marcelo Finger, Maria das Graças V. Nunes and Stella E. Tagnin (2003). The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In Proceedings of Corpus Linguistics 2003 Conference (CL2003), Lancaster, UK.

Archer, Dawn, Paul Rayson, Scott Piao, Tony McEnery (2004). Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies. In Williams G. and Vessier S. (eds.) Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004), Lorient, France. Volume III, pp. 817-827.

Balossi, Giuseppina (2014) A Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's The Waves. Benjamins.

Borghetti, Claudia, Sara Castagnoli and Marco Brunello (2011). I testi del web: una proposta di classificazione sulla base del corpus PAISÀ. In Cerruti, M., E. Corino and C. Onesti (eds.): Scritto e parlato, formale e informale: La comunicazione mediata dalla rete., Roma: Carocci, pp. 147-170.

Brooke, Julian, Milan Tofiloski, and Maite Taboada (2009). Cross-linguistic sentiment analysis: From English to Spanish. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), pp. 50-54.

Chitchyan, Ruzanna, Américo Sampaio, Awais Rashid and Paul Rayson (2006). Evaluating EA-Miner: Are Early Aspect Mining Techniques Effective? In proceedings of Towards Evaluation of Aspect Mining (TEAM 2006). Workshop Co-located with ECOOP 2006, European Conference on Object-Oriented Programming, 20th edition, Nantes, France, pp. 5-8.

Cunningham, Hamish, Diana Maynard, Kalina Bontcheva (2011). Text Processing with GATE. Gateway Press CA. ISBN: 0956599311 9780956599315.

Davies, Mark and Ana Preto-Bay (2007). A Frequency Dictionary of Portuguese. Routledge. ISBN-10: 0415419972.

Doherty, Neil, Nigel Lockett, Paul Rayson and Stuart Riley (2006). Electronic-CRM: a simple sales tool or facilitator of relationship marketing? The 29th Institute for Small Business & Entrepreneurship Conference. International Entrepreneurship - from local to global enterprise creation and development, Cardiff-Caerdydd, UK.

Gacitua, Ricardo, Pete Sawyer, Paul Rayson (2008). A flexible framework to experiment with ontology learning techniques. In Knowledge-Based Systems, 21(3), pp. 192-199.

Hancock, Jeffrey, T., Michael T. Woodworth and Stephen Porter (2013) Hungry like the wolf: A word-pattern analysis of the language of psychopaths. Legal and Criminological Psychology. 18 (1) pp. 102-114.

Hermann, Karl Moritz and Phil Blunsom (2013). A Simple Model for Learning Multilingual Compositional Semantics. arXiv:1312.6173 [cs.CL]. URL: http://arxiv.org/abs/1312.6173.

Klebanov, Beigman B., Daniel Diermeier and Eyal Beigman (2008). Automatic annotation of semantic fields for political science research. Journal of Language Technology and Politics 5(1), pp. 95-120.

Löfberg, Laura, Scott Piao, Asko Nykanen, Krista Varantola, Paul Rayson, and Jukka-Pekka Juntunen (2005). A semantic tagger for the Finnish language. In the Proceedings of the Corpus Linguistics Conference 2005, Birmingham, UK.

McArthur, Tom (1981). Longman Lexicon of Contemporary English. Longman London.

McEnery, Tony and Zhonghua. Xiao (2004). The Lancaster corpus of Mandarin Chinese: a corpus for monolingual and contrastive language study. In Proceedings of LREC 2004. pp. 1175-1178.

Markowitz DM, Hancock JT (2014) Linguistic Traces of a Scientific Fraud: The Case of Diederik Stapel. PLoS ONE 9(8): e105937.

Mudraya, Olga, Bogdan Babych, Scott Piao, Paul Rayson and Andrew Wilson (2006). Developing a Russian semantic tagger for automatic semantic annotation. In Proceedings of the International Conference "Corpus Linguistics - 2006", St.-Petersburg, Russia, pp. 290-297.

Nakano, Tomofumi and Yukie Koyama (2005). e-Learning Materials Development Based on Abstract Analysis Using Web Tools. Knowledge-Based Intelligent Information and Engineering Systems. In Proceedings of KES 2005, Melbourne, Australia, LNCS 3681, Springer, pp. 794-800. DOI 10.1007/11552413_113.

Nasiruddin, Mohammad (2013). A State of the Art of Word Sense Induction: A Way Towards Word Sense Disambiguation for Under-Resourced Languages. arXiv:1310.1425 [cs.CL]. URL: http://arxiv.org/abs/1310.1425.

Ooi, Vincent B.Y., Peter K.W. Tan and Andy K. L. Chiang (2007). Analyzing personal weblogs in Singapore English: the Wmatrix approach. Studies in Variation, Contacts and Change in English. Volume 2. Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki.

Padró, Lluís and Evgeny Stanilovsky (2012). FreeLing 3.0: Towards Wider Multilinguality. In Proceedings of the Language Resources and Evaluation Conference (LREC 2012). Istanbul, Turkey. May, 2012.

Popov, Borislav, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff and Miroslav

Goranov (2003). KIM - Semantic Annotation Platform. In Proceedings of 2nd International Semantic Web Conference (ISWC2003), Florida, USA, pp. 834-849.

Potts, Amanda and Paul Baker (2013). Does semantic tagging identify cultural change in British and American English?, International Journal of Corpus Linguistics 17(3): 295-324.

Rayson, Paul, Dawn Archer, Scott Piao, Tony McEnery (2004). The UCREL semantic analysis system. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, pp. 7-12.

Reeve, Lawrence and Hyoil Han (2005). Survey of Semantic Annotation Platforms. Proceedings of the 2005 ACM Symposium on Applied Computing, pp. 1634—1638.

Schmid, Helmut (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Sharoff, Serge. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni and S. Bernardini (eds.), WaCky! Working papers on the Web as Corpus. Bologna, Italy: Gedit.

Taiani, Francois, Paul Grace, Geoff Coulson and Gordon Blair (2008). Past and future of reflective middleware: Towards a corpus-based impact analysis. The 7th Workshop On Adaptive And Reflective Middleware (ARM'08) December 1st 2008, Leuven, Belgium, collocated with Middleware 2008.

Toutanova, Kristina, Dan Klein, Christopher Manning, and Yoram Singer (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.

Volk, Martin, Bärbel Ripplinger, Spela Vintar, Paul Buitelaar , Diana Raileanu , Bogdan Sacaleanu (2002). Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval. International Journal of Medical Informatics 67(1-3), pp. 97-112.

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

Xiao, Richard, Paul Rayson and Tony McEnery (2009). A Frequency Dictionary of Mandarin Chinese: Core Vocabulary for Learners. Routledge. ISBN-10: 0415455863.

Zhang, Lei and Achim Rettinger (2014). Semantic Annotation, Analysis and Comparison: A Multilingual and Cross-lingual Text Analytics Toolkit. In Proceedings of the Demonstrations at the EACL 2014, Gothenburg, Sweden, pp. 13-16.