Tools for plWordNet Development: Presentation and Perspectives

Bartosz Broda, Marek Maziarz, Maciej Piasecki

Institute of Informatics, Wrocław University of Technology,
{bartosz.broda, marek.maziarz, maciej.piasecki }@pwr.wroc.pl

Abstract

Building a wordnet is a serious undertaking. Fortunately, Language Technology (LT) can improve the process of wordnet construction both in terms of quality and cost. In this paper we present LT tools used during the construction of *plWordNet* and their influence on the lexicographer's work-flow. LT is employed in *plWordNet* development on every possible step: from data gathering through data analysis to data presentation. Nevertheless, every decision requires input from the lexicographer, but the quality of supporting tools is an important factor. Thus a limited evaluation of usefulness of employed tools is carried out on the basis of questionnaires.

Keywords: wordnet, *plWordNet*, computer-aided lexicography, wordnet expansion

1. Introduction

A wordnet is a lexico-semantic network build in a spirit of Princeton WordNet (Fellbaum, 1998). Usually, *lexical units* (word senses) in wordnets are organised in *synets*, i.e., a set of near synonyms. Both lexical units and synsets are interconnected with each others via lexico-semantic *relations*. A typical synset realtion is a hyperonymy/hyponymy, which express *is a kind of* relation, e.g., a dog is a kind of an animal. There were many initiatives aimed at building wordnets for many languages, including both international efforts (EuroWordNet (Vossen, 2002), BalkaNet (Tufiş et al., 2004), Global WordNet Grid, e.g. (Vossen et al., 2008)) and projects limited only to national level (e.g., Princeton WordNet, sloWNet (Fišer and Sagot, 2008), WOLF (Sagot and Fišer, 2012)).

A wordnet is very useful in many Natural Language Processing (NLP) tasks. For example, Information Extraction (IE), Machine Translation (MT) or Question Answering (QA) all benefit from wordnet availability. Thus a language without a wordnet is at sever disadvantage. How should one proceed if there is no wordnet for hers/his language?

Building a wordnet is a serious undertaking. Process of wordnet construction can be perceived as a dictionary building task, e.g. (Fellbaum, 1998). Lexicographers distinguish several phases in preparing a dictionary. From data collection and selection through data analysis to data presentation are all hard and time consuming tasks (Svensén, 2009). With the advent of personal computers the lexicography practices were revolutionised, but the above steps remain largely unchanged. Nevertheless, the usage of Language Technology (LT) can improve the process of wordnet construction both in terms of efficiency and coverage. In this paper we would like to present how LT can be utilised for this purpose. We want to share experiences gained during the construction of Polish wordnet called *plWordNet*. The process of *plWordNet* expansion, tools involved in it, and their envisaged further research and development are discussed.

plWordNet 1.0 emerged in 2009 and since then its development has been continued. Till now (i.e., 14.03.2012) it has reached the size of: 89,291 lemmas, 135,400 lexical units and 96,644 synsets. At the very beginning we assumed that translation from another language (e.g., Princeton WordNet) is inappropriate, because we wanted to obtain as faithful description of the Polish lexical system as possible. We also could not base our work on transferring knowledge from any existing electronic dictionary. Instead we decided to apply a bottom-up, corpus-driven approach. This methodology imposed the need of developing and utilizing computer-aided approach as we had a limited budged and limited time for *plWordNet* construction. We assumed that the constructed tools should encompass as wide spectrum of wordnet-editing phases as possible. It is a well known problem that editing of such a large thesaurus by multiple people is a difficult task and constant assistance of software can improve the process.

LT is a driving force behind the development of tools supporting linguist' work. The algorithms employed typically can be automatically evaluated. On the other hand, evaluation of *usefulness* of tools and their contribution to the whole process is not as straightforward. Thus we decided to perform limited usability evaluation on the basis of questionnaire. After several months of *plWordNet* expansion each linguist had to answer some questions about their work. Eight linguists took part in the survey. Although the sample is rather small (but it consists of the whole population of lexicographers working on *plWordNet*), we still gained some interesting insights about the LT employed. We will discuss those findings during the description of tools.

2. Data collection and selection

The first phase in constructing a new dictionary is collecting fundamental resources, i.e., corpora (Sinclair, 2003). Next, some data should be selected from the vast amount of linguistic material for further analysis. We used **three corpora** (henceforth, the *joint corpus*): 250 million token ICS PAS Corpus (Przepiórkowski, 2004), 113 million token corpus of texts from "Rzeczpospolita" (Rze, 1993 2002) and a large corpus of texts collected from the Internet (ca 800 million tokens). Every stage of our wordnet building process is corpus-dependent. Due to the requirements of several tools used in the process, the texts must be first tagged morpho-syntactically: a morphological analyser *Morfeusz* (Woliński, 2006) and TaKIPI tagger (Piasecki, 2007) were applied for Polish.

After initial step of gathering and preprocessing of corpora we have to select lemmas for inclusion in *plWordNet*. We base the lemma-selection process on the frequency dictionary extracted from the joint corpus. Naturally, we start with the most frequent lemmas.

However, the frequency criterion should not limit *plWordNet* editors in broadening the list with additional lemmas that are very common in spoken Polish. Also common words that do not occur in the joint corpus, but are obtained by introspection or noticed in dictionaries can extend the initial frequency list. The former reason is motivated that we do not use a reliable corpus of spoken Polish thus this register of language is under-represented on the frequency list. The latter reason is motivated by the fact that no corpus is perfectly balanced (esp. when large parts of corpus are taken from the Internet) and even the biggest corpus will not reflect the language perfectly.

In this phase a special attention should be paid to multiword expressions (MWEs), i.e. multiword lemmas. Discovering MWEs received a lot of attention in literature, e.g., (Evert, 2004). However, in the case of an inflectional language application of statistical measures of association to lemma sequences in the lemmatised corpus, performed with help of Kolokacje (Buczyński, 2004) tool, was only the first step. Next, the approach was extended with automated identification of statistically significant morpho-syntactic patterns for MWEs (Broda et al., 2008).

It would not be feasible to build a dictionary without consulting other dictionaries, so the next step in dictionary making is choosing existing dictionaries as reference (Svensén, 2009, p. 428). It is problem-prone sphere, because of legal and ethical aspects (Svensén, 2009, ch. 25). We make use of several modern dictionaries, e.g., (Dubisz, 2004), (Bańko, 2000), (SJP, 2011), (MSS, 2011), popular encyclopaedias, i.e., *Wikipedia*, (Enc, 2011), and many thematic lexicons but only as secondary and supplementary sources. The main idea of building a wordnet as close to the corpus as possible has been kept unchanged throughout the *plWordNet* construction process.

3. Data analysis and presentation

During the data analysis phase lexicographers are responsible for:

- deciding whether lemmas presented to them are real units of Polish,
- distinguishing lemma senses, which is a very difficult task,
- attaching new senses (lexical units) to the wordnet structure.

Such a data analysis would be hardly feasible when performed on a huge corpus without LT. As a part of *plWordNet* project several tools were developed with the aim to support wordnet editing and lexico-semantic corpus analysis.

3.1. plWordNet Application and WordnetLoom

The core system for the whole *plWordNet* development process is a wordnet editor which was called initially *plWordNetApp* (Piasecki and Koczan, 2007), and recently expanded to *WordnetLoom* (Marcińczuk and Piasecki, 2011b). *plWordNetApp*, which originated from the manual construction of the core of *plWordNet*, enabled network-based collaborative editing of *plWordNet*. The GUI-based editing environment was designed to supplement the developed process of *plWordNet* construction, e.g., via automatically generated substitution tests for relations.

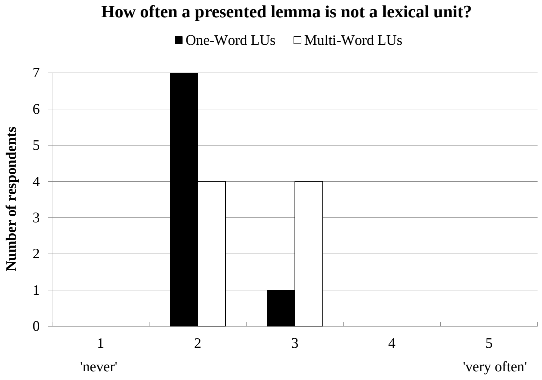

Figure 1: Accuracy of lemma recognition according to *plWordNet* editors.

WordnetLoom is a graph-based wordnet editor enabling visual browsing and editing the *plWordNet* structure. It is also closely integrated with semi-automatic tools for wordnet expansion. Since deployment of *WordnetLoom* in the beginning of 2011 linguists have been encouraged to work in graph-based visualisation perspective. Now most of them use this tool almost exclusively (5 of 8 persons answered that they use *WordnetLoom* as a main *plWordNet* editor).

3.2. Primary and secondary linguistic sources

Lemma verification is a simple task performed on the basis of linguistic intuition, corpus browsing and secondary sources (i.e., query in dictionaries, encyclopaedias and lexicons). Our automatic methods achieved quite good results in the questionnaire, for question "How often a presented lemma is not a lexical unit?" (Fig. 1) most of editors gave the answer '2' (we interpret it as 'rare', Fig. 1). It is not surprising that MWEs recognition performs a little bit worse. *plWordNet*lexicographers claimed that the most common mistakes of automatic methods for MWE were: an inappropriate word order and an inappropriate MWE base form.

While editing the wordnet structure linguist needs to define lemma senses (lexical units) and constraints (expressed by the relation network) between them. Determination of senses boundaries is a very difficult problem. Sometimes intuition about distinctions between different word senses for a given lemma can be misleading. As we settled on corpus-based approach to wordnet construction, we can encourage lexicographers to not to depend exclusively on intuition. Linguists are expected to check usages of lemmas in the available corpora using Poliqarp interface (Janus and Przepiórkowski, 2006). They may use also Internet (e.g., Google) to search Web collections. The resources are of high usability in lexicographers work (Fig. 2). Neverthe-

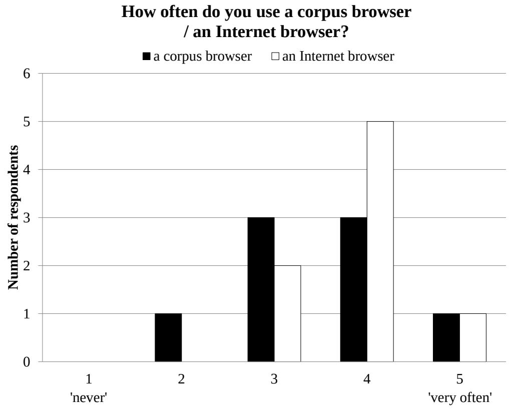

Figure 2: A corpus browser and an Internet browser usability comparison.

less, the usability comes at the cost of the effort required to utilise them (e.g., finding examples of rare senses of words in a large corpus is difficult and time consuming).

3.3. Examples of usages generated by a WSD algorithm

In order to speed-up the process of finding different word senses we employ an improved version of the LexCSD algorithm (Broda et al., 2010). LexCSD preforms automatic clustering of lemma occurrences based on their distributional usage patterns. The clustering algorithm is automatically fine-tuned to the data and the number of word senses is estimated on the basis of corpora. After forming the clusters a few most representative *usage examples* are selected by the algorithm and presented to the linguists in WordNetLoom. The selection of examples uses multi-criteria heuristic approach. We take into account centrality of an example in relation to other examples in the cluster, proportion of part of speeches in the example, named entities (Marcińczuk and Piasecki, 2011a), hand written rules, etc. At the end of the process the linguists are presented with a few usage examples for different senses of a given lemma.

Linguists appraised this tool highly. Having been asked the question "How often do you use a corpus browser / WSD examples of usages?" linguists marked higher grades for WSD examples than for corpus browsers (Fig. 3). The result is in agreement with answers for the question "What is less time-consuming: working with WSD examples or with corpus browser?" (Fig. 4). *plWordNet* editors answered unanimously that working with WSD examples of usages helped saving time. Fig. 3 shows one more interesting pattern: usability of WSD examples and dictionaries are comparable, this proves that the WSD tool performs very well.

3.4. WordnetWeaver

Sense distinguishing has been also supported during wordnet expansion process by the WordnetWeaver system (with user interface integrated with WordnetLoom as one of its screens). Its main task is to generate and present to the linguists suggested lexical units (senses) for a set of new lemmas (not yet described in *plWordNet*). Each suggestion for a new lemma *x* is presented on the screen in a form of hy-

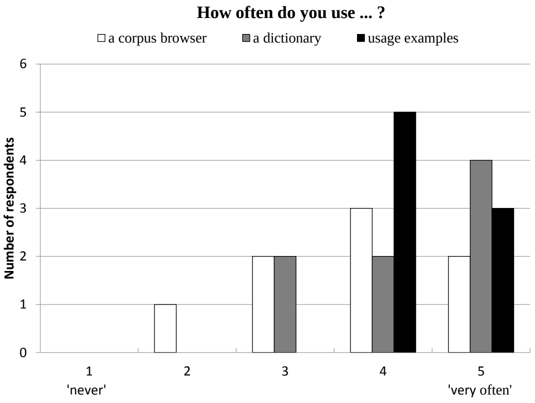

Figure 3: Disambiguated usage examples, corpus browser and dictionary usability comparison.

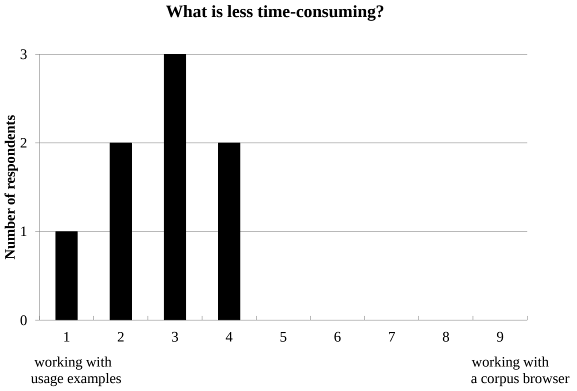

Figure 4: Comparison of estimated effort while working with disambiguated usage examples and a corpus browser.

pernymic sub-graph. Each synset in a sub-graph expresses a significant *semantic fit* to *x* (Piasecki et al., 2009).

The semantic fit is based on heterogeneous knowledge sources describing lemma-to-lemma semantic associations that are extracted from a large corpus. The applied methods encompass: methods of Distributional Semantics – measures of semantic relatedness, pattern-based methods (manual and automated) and Machine Learning (i.e., a classifier for lemma pairs as belonging to a wordnet relation). Measures of semantic relatedness are extracted with the help of SuperMatrix system implementing many methods of Distributional Semantics (Broda and Piasecki, 2008).

Semantic fit between *x* and a synset (or a lexical unit in the recent, extended version of the algorithm (Piasecki et al., 2011)) is based not only on knowledge sources describing association between *x* and the given synset members (or lexical unit lemma) but also on the basis of the local context of the synset in the wordnet relation graph (with emphasis given to the hypernymy graph).

WordnetWeaver algorithm expresses high precision for the top ranked suggestions (i.e., sub-graphs) (Broda et al., 2011), but there is no natural delimitation of reliable suggestions in terms of the semantic fit level. As a result, up to *k* suggestions are presented for each lemma, where *k* is

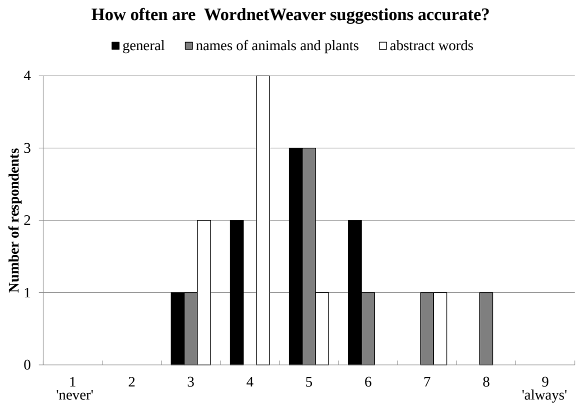

Figure 5: Subjective WNW accuracy and accuracy in specific domains of concrete words (names of plants and animals) and abstract words.

set as a parameter of the algorithm. Thus, non-relevant suggestions can blur the picture in some cases. Moreover, in spite of using heterogeneous knowledge extraction methods the whole approach is limited by the lemma senses significantly supported by the corpus evidence. Measures of semantic relatedness are typically biased towards most frequent senses of words in the corpus. Still, WordnetWeaver automates wordnet expansion in the case of more concrete nouns, helps to locate quickly wordnet sub-graphs that are relevant to a new lemma and draws linguists' attention to senses specific to the corpus.

WordnetWeaver offers a kind of high level, semantic corpus browsing integrating heterogeneous evidence for semantic lemma associations. Its main advantage over concordance-based corpus browsing or disambiguated examples is that it shows a direct place in the wordnet graph where the link between new and existing lexical units could be established. This is also sometimes its main drawback as for rare word senses the linguist have to analyse the network structure in order to understand the suggestion of algorithm. *plWordNet* lexicographers gave positive assessments of WordnetWeaver accuracy (5). It is interesting that – according to *plWordNet* editors – WNW acts better when applied to concrete nouns (names of plants and animals) than to abstract words – this is probably connected with more sophisticated structure of concrete noun lexical fields. However, the high precision of the tool effects completeness which is lower than possible, this weakens a little usability of WNW in comparison with dictionaries or WSD examples (Fig. 5). Thus, it seems that WordnetWeaver and recently added disambiguated examples from the corpus complement each other perfectly.

3.5. Automatic Recognition of Derivational Relations

Derivational relations (based on word form dependencies) are present in many languages and described in many wordnets due to their often systematic association with semantic oppositions. However, they are significantly more numerous and important as a part of the lexical system in Slavic languages, including Polish.

A language tool generating derivational pairs (e.g., an expanded morphological analyser) is required to automate wordnet expansion in this domain. Such a tool doest not exist for Polish and many other languages too. Instead, an analyser of derivational pairs, called *Derivator*, which is based on transducers trained on pairs already described in the wordnet and extended with automated construction of mappings representing internal stem alternations was built. It can be easily adapted to other languages. However, as derivational relations hold for specific lexical units in *plWordNet*, pairs recognised by Derivator as matching word form level patterns for the particular derivational relations must be additionally filtered due to their semantic properties in order to obtain practical accuracy. The properties should be extracted from the corpus. The overall solution is under development now.

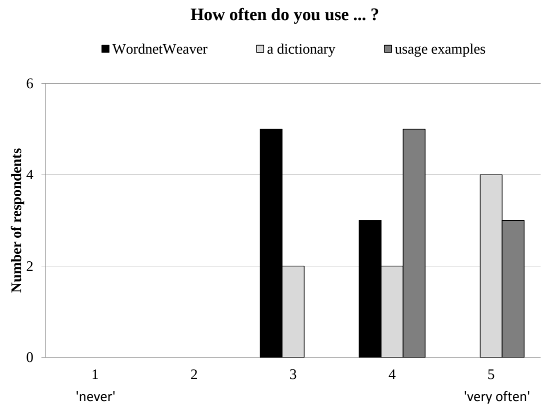

Figure 6: Subjective WordnetWeaver usability.

4. Perspectives

Above we have briefly described the process of transforming the joint corpus to a lemma list and beyond (from data collection to data presentation). Several tools were used in phases of data collection and selection: morphological analyser (*Morfeusz*), tagger (TaKIPI), tools for MWE recognition. During the data analysis stage linguists were equipped with four applications: the joint corpus browser (Poliqarp), disambiguated examples of usage (based on LexCSD algorithm), WordnetWeaver system (including the SuperMatrix) and Derivator. All the work – editing and presentation – is done in WordnetLoom environment.

Our next aim is to extend the WordnetWeaver with a form of word sense disambiguation and integrate it with automatically retrieved sense-usage examples. We want to semi-automatically extend *plWordNet* with those sense-usage examples. There is no representative dictionary without a representative corpus: we constantly expand our web corpus in order to diversify genres and stylistic registers of available texts during the text processing phases.

Acknowledgements

Work financed by Innovative Economy Programme project POIG.01.01.02-14-013/09 (`http://www.ipipan.waw.pl/nekst/`).

5. References

Mirosław Bańko, editor. 2000. *Inny słownik języka polskiego PWN*, volume 1-2. Wydawnictwo Naukowe PWN, Warszawa.

Bartosz Broda and Maciej Piasecki. 2008. SuperMatrix: a general tool for lexical semantic knowledge acquisition. In Grażyna Demenko, Krzysztof Jassem, and Stanisław Szpakowicz, editors, *Speech and Language Technology*, volume 11, pages 239–254. Polish Phonetics Assocation. Tthe first version was published in the Proceedings of the International Multiconference on Computer Science and Information Technology — 3rd International Symposium Advances in Artificial Intelligence and Applications (AAIA'08).

B. Broda, M. Derwojedowa, and M. Piasecki. 2008. Recognition of structured collocations in an inflective language. *Systems Science*, 34(4):27–36. extended version of a paper in the Proceedings of AAIA'08, Wisła, Poland.

B. Broda, M. Maziarz, and M. Piasecki. 2010. Evaluating lexcsd — a weakly-supervised method on improved semantically annotated corpus in a large scale experiment. In S. T. Wierzchoń M. A. Kłopotek, A. Przepiórkowski and K. Trojanowski, editors, *Proceedings of Intelligent Information Systems*.

Bartosz Broda, Roman Kurc, Maciej Piasecki, and Radosław Ramocki. 2011. Evaluation method for automated wordnet expansion. In P. Bouvry, M. Kłopotek, F. Leprevost, M. Marciniak, A. Mykowiecka, and H. Rybiński, editors, *Security and Intelligent Information Systems*, LNCS. Springer. To appear.

A. Buczyński. 2004. Pozyskiwanie z internetu tekstów do badań lingwistycznych.

Stanisław Dubisz, editor. 2004. *Uniwersalny słownik języka polskiego [a universal dictionary of Polish], electronic version 1.0.* PWN.

2011. Encyklopedia pwn.

S. Evert. 2004. The statistics of word cooccurrences: word pairs and collocations. *Unpublished doctoral dissertation, Institut f
"ur maschinelle Sprachverarbeitung, Universit
"at Stuttgart.*

Christiane Fellbaum, editor, 1998. *WordNet. An Electronic Lexical Database*, chapter Introduction. The MIT Press.

Darja Fišer and Benoít Sagot. 2008. Combining multiple resources to build reliable wordnets. In *Text, Speech and Dialogue*, number 2546 in LNCS, pages 61–68. Springer, Berlin; Heidelberg.

Daniel Janus and Adam Przepiórkowski. 2006. Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora. In Jacek Waliński, Krzysztof Kredens, and Stanisław Goźdź-Roszkowski, editors, *The proceedings of Practical Applications of Linguistic Corpora 2005*. Peter Lang.

Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors. 2006. *Intelligent Information Processing and Web Mining – Proceedings of the International IIS: IIPWM '06 Conference held in Wisła, Poland, June, 2006.* Advances in Soft Computing. Springer, Berlin.

Michał Marcińczuk and Maciej Piasecki. 2011a. Statistical Proper Name Recognition in Polish Economic Texts. *Control and Cybernetics*. To Appear.

Michał Marcińczuk and Maciej Piasecki. 2011b. Wordnet-Loom: a wordnet development system integrating form-based and graph-based perspectives. *International Journal of Data Mining, Modelling and Management*. Accepted for publishing.

2011. Miejski słownik slangu i mowy potocznej.

Maciej Piasecki and Paweł Koczan. 2007. Environment supporting construction of the Polish Wordnet. In Zygmunt Vetulani, editor, *Proceedings of the 3rd Language and Technology Conference, October 5–7, 2007, Poznań, Poland*, pages 519–523, Poznań. Wydawnictwo Poznańskie Sp. z o.o.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

Maciej Piasecki, Roman Kurc, and Bartosz Broda. 2011. Heterogeneous knowledge sources in graph-based expansion of the polish wordnet. In *Proceedings of The 2nd Asian Conference on Intelligent Information and Database Systems*, number 6591 in LNAI. Springer. To appear.

Maciej Piasecki. 2007. Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, 11(1–2):151–167.

Adam Przepiórkowski. 2004. *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

1993-2002. Korpus rzeczpospolitej. [on-line] `http://www.cs.put.poznan.pl/dweiss/rzeczpospolita`.

Benoît Sagot and Darja Fišer. 2012. Automatic extension of WOLF. In *Proceedings of the 6th International Global Wordnet Conference, Matsue, Japan*, pages 317–323. The Global WordNet Association.

John Sinclair. 2003. Corpora for lexicography. In Piet van Sterkenburg, editor, *A Practical Guide to Lexicography*, pages 167–214. John Benjamins Publishing Co.

2011. Słownik języka polskiego pwn.

Bo Svensén. 2009. *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge University Press.

Dan Tufiş, Dan Cristea, and Sofia Stamou. 2004. Balka-Net: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information Science and Technology*, 7(1–2):9–43. Special Issue.

Piek Vossen, Eneko Agirre, Nicoletta Calzolari, Christiane Fellbaum, Shu-Kai Hsieh, Chu-Ren Huang, Hitoshi Isahara, Kyoko Kanzaki, Andrea Marchetti, Monica Monachini, Federico Neri, Remo Raffaelli, German Rigau, Maurizio Tesconi, and Joop VanGent. 2008. KYOTO: A system for mining, structuring, and distributing knowledge across languages and cultures. In A. Tanács, D. Csendes, V. Vincze, Ch. Fellbaum, and P. Vossen, editors, *Proceedings of the Fourth Global WordNet Conference*, pages 474–484.

Piek Vossen. 2002. EuroWordNet general document version 3. Technical report, University of Amsterdam.

Marcin Woliński. 2006. Morfeusz – a practical tool for the morphological analysis of Polish. In Kłopotek et al. (Kłopotek et al., 2006), pages 511–520.