# A Survey of WordNet Annotated Corpora

**Tommaso Petrolito**[⊕⊙] and **Francis Bond**[⊕]
[⊕]Linguistics and Multilingual Studies,
**Nanyang Technological University**, Singapore
[⊙] Informatica Umanistica,
**University of Pisa**, Italy
bond@ieee.org, tommasouni@gmail.com

## Abstract

This paper surveys the current state of wordnet sense annotated corpora. We look at corpora in any language, and describe them in terms of accessibility and usefulness. We finally discuss possibilities in increasing the interoperability of the corpora, especially across languages.

## 1 Introduction

There are over 60 different wordnet projects projects for more than 60 languages.[1] The first wordnet was the Princeton WordNet of English (Fellbaum, 1998) describing over 150,000 concepts. Many others have followed, even if with different coverage rates in each continent (Africa and central Asia are less covered than the other geographical regions), all around the world. So today there are many wordnets all sharing a similar structure, some of them freely available, others restricted to license owners.

Bond and Paik (2012) surveyed the available wordnets and evaluated them on two axes: how **accessible** (legally OK to use) and how **usable** (of sufficient quality, size and with a documented interface) (Ishida, 2006). In this paper we do the same for sense-annotated corpora. We restrict ourselves to those that use a wordnet as the sense inventory.

Sense annotated corpora can be classified according to several criteria. Some obvious ones are the language used; the lexicon used to determine the senses; the size; the license. In addition, another useful distinction is that between those that annotate **all words** and those that only annotate **some words**, typically either a sample of a few frequent words, or of a single part-of-speech. We will also distinguish those corpora that align to SemCor (Langone et al., 2004) the first wordnet annotated corpus. We will first describe it in some detail, as it is the most typical corpus, and then note where other corpora differ from it.

We have found more than 20 WordNet Annotated Corpora in more than 10 different languages. We describe them in the following Section 2, discuss some of the issues they raise in Section 3 and then plans for future work in 4.

[1] `http://globalwordnet.org/?page_id=38`

## 2 WordNet Annotated Corpora

We have tried to list all known corpora annotated with wordnet senses, in any language.[2] In most cases, information on size comes from the latest publication describing the corpus, or its web-page. Sometimes the data is from the corpus providers themselves, in which case we will note this. We have also put the information online as the Global Wordnet Association's *Wordnet Annotated Corpora* page (`http://globalwordnet.org/?page_id=241`). This will be kept up-to-date.

We divide the corpora into three groups: SemCor and its translations; non-English Corpora; and English Corpora. We summarize the corpora in Table 1, and then describe each one in more detail.

### 2.1 SemCor and Translations

#### 2.1.1 Princeton SemCor

The English SemCor corpus is a sense-tagged corpus of English created at Princeton University by the WordNet Project research team (Landes et al., 1998). It was created very early in the WordNet project (Miller et al., 1994), and was one of the first sense-tagged corpora produced for any language. The corpus consists of a subset of the Brown Corpus (700,000 words, with more than 200,000 sense-annotated) (Francis and Kucera, 1979), and it has been part-of-speech-tagged and sense-tagged. It is distributed under the Princeton Wordnet License.

For each sentence, open class words (or multi-word expressions) and named entities are tagged. Not all expressions are tagged. We give a (constructed) example in Figure 1. Note that the tagged synsets do not have to be continuous (as in *get up*) and that there are some untagged elements (typically multi word expressions, such as ***on one's feet***). Closed class words such as articles and prepositions are only tagged if they are part of a multi-word expression. The annotation is known to be imperfect: Bentivogli and Pianta (2005) estimate around 2.5% of the tags to be incorrect.

The Brown corpus has also been annotated with syntactic information by various other projects, including the Penn Treebank (Marcus et al., 1993); Susanne (Sampson, 1995) (also sense-annotated with the WordNet 1.6 senses in the SemiSusanne project by Powell (2005)) and Redwoods (Oepen et al., 2004; Flickinger,

[2] Although we may have missed some lexical sample corpora.

| Name | # words | # taggable | # tagged | lng | Wordnet | License | Semcor | Target |
|---|---|---|---|---|---|---|---|---|
| SemCor3.0-all | 360k | n/a | 193k | eng | WN 3.0 | wordnet | + | all |
| SemCor3.0-verbs | 317k | n/a | 41k | eng | WN 3.0 | wordnet | + | v |
| Jsemcor | 380k | 150k | 58k | jpn | Jpn WN | wordnet | + | all |
| MultiSemCor[a] | 269k | 121k | 93k | ita | MultiWN | CC BY 3.0 | + | all |
| | 258k | n/a | 120k | eng | WN 1.6 | CC BY 3.0 | | |
| SemCor EnRo | 176k | 89k | 48k | rum | BalkaNet | MSC ... | + | all |
| | 178k | n/a | n/a | eng | WN 2.0 | BY-NC-ND | | |
| BulSemCor[b] | 101k | n/a | 99k | bul | BulNet | web only | – | all+ |
| Eusemcor | 300k | n/a | n/a | baq | Basque WN | web only | – | all |
| spsemcor | 850k | n/a | 23k | spa | ESPWN1.6 | web only | – | n, v |
| AnCora | 500k | n/a | n/a | spa | EuroWN 1.6 | research only | – | n |
| | 500k | n/a | n/a | cat | EuroWN 1.6 | research only | | |
| DutchSemcor[c] | 500,000k | n/a | 283k | dut | Cornetto | n/a | – | all |
| TüBa-D/Z Treebank[d] | 1,365k | n/a | 18k | ger | GermaNet | none | – | some, v, n |
| WebCaGe | n/a | n/a | 11k | eng | GermaNet | CC BY-SA 3.0 | – | all |
| ISST | 306k | n/a | 81k | ita | ItalWN | research only | – | all |
| NTU-MC | 116k | 63k | 51k | eng | PWN | CC BY | – | all |
| | 106k | 67k | 36k | cmn | COW | CC BY | | |
| | 56k | 37k | 28k | ind | WN Bahasa | CC BY | | |
| | 49k | 20k | 15k | jpn | Jpn WN | CC BY | | |
| AQMAR Arabic SST[e] | 65k | n/a | 32k | ara | WN | CC BY-SA 3.0 | – | n, v |
| Jos100k[f] | 100k | n/a | 5k | slv | sloWNet | CC BY-NC 3.0 | – | some n |
| Hungarian WSD corpus | 16k | n/a | 5k | hun | HuWN | none | – | n, v, adj |
| KPWr | 438k | n/a | 9k | pol | plwordnet | CC BY 3.0 | – | some |
| Gloss Corpus | 1,621k | 656k | 449k | eng | WN 3.0 | wordnet | – | some |
| Groningen Meaning Bank | 1,020k | n/a | n/a | eng | WN | none | – | all |
| MASC | 504k | n/a | 100k | eng | WN 3.0 | none | – | v |
| DSO Corpus | n/a | n/a | 193k | eng | WN 1.5 | LDC | – | n, v |
| OntoNotes | 1,500k | n/a | n/a | eng | Coarse WN | LDC | – | n, v |
| SemLink | 78k | n/a | n/a | eng | Coarse WN | none | – | all |
| Senseval 3 | 5k | n/a | 2k | eng | WN 1.7.1 | none | – | all |
| SemEval-2013 Task 12[g] | 5k | n/a | n/a | eng | BabelNet | none | – | n |
| SemEval-2013 Task 13 | 141k | n/a | 5k | eng | BabelNet | none | – | n, v, adj |

Table 1: Corpora Tagged with Wordnet Senses

**a** According to Bentivogli and Pianta (2005) 23.4% of Italian words still need to be tagged, so we can estimate (given that 93k is the 76.6%) the content words at 121k.

**b** The annotations include both open-class and closed-class words.

**c** 282,503 tokens manually tagged by two annotators, anyway more than 400,000 have been manually tagged by at least one annotator and millions have been automatically tagged (information from the corpus providers themselves: Piek Vossen).

**d** The targets of the annotation are not all the nouns and verbs but only a selected set of 109 words (30 nouns and 79 verbs). The total number of annotations is 17,910 (information from the corpus providers themselves: Verena Henrich and Marie Hinrichs). The corpus is not currently available but it will be.

**e** According to Schneider et al. (2012) about half the tokens in the data are covered by a nominal supersense, so we can estimate (given that the tokens are 65k) the tagged tokens at 32k.

**f** Only the 100 most frequent nouns are annotated.

**g** The corpus is multilingual, in fact the same articles are available in other four languages:french, spanish, german and italian, respectively containing 3k tokens each, Frech, Spanish and German and 4k Italian)

*Kim$_a$ got$_b$ slowly$_c$ up$_b$, the children$_d$ were$_e$ already$_f$ on$_g$ their$_g$ feet$_g$.*

| ID | Lemma | Sense |
|---|---|---|
| a | Kim | **org** |
| b | get_up | get_up$_4$ |
| c | slowly | slowly$_1$ |
| d | child | child$_1$ |
| e | be | be$_3$ |
| f | already | already$_1$ |
| g | on_one's_feet | **notag** |

Figure 1: SemCor Example

2011). The combination of syntactic and semantic information has been used in various parsing experiments (Bikel, 2000; Agirre et al., 2008). The corpus is divided into two parts: **semcor-all** in which 186 texts have all open-class words (such as nouns, verbs, adjectives and adverbs) semantically annotated. The SemCor component of all word types consists of 359,732 (Lupu et al., 2005) tokens of which 192,639 are semantically annotated. The second part, **semcor-verbs**, only has verbs senses annotated: 41,497 verbal occurrences from 316,814 tokens (Lupu et al., 2005).

#### 2.1.2 MultiSemCor

MultiSemCor is an English/Italian parallel corpus created by translating the English SemCor corpus into Italian (Bentivogli and Pianta, 2005). In particular it consisted of the translation of 72% of the SemCor-all corpus. This sub-corpus was automatically word aligned and the semantic annotations were automatically projected from the English words to their Italian translation equivalents. The resulting corpus has texts aligned both at the sentence and word level, and annotated with part of speech, lemma and word sense (PWN 1.6). MultiSemCor version 1.1 contains 14,144 sentences and 261,283 tokens, 119,802 of which are annotated with senses. Words that did not project from English were not tagged: an estimated 23.4% of the concepts that should be tagged are not. The MultiSemCor project includes a MultiSemCor Web Interface (Ranieri et al., 2004). It provides for two distinct browsing modalities. In the *text-oriented* modality (*MSC Browser*), for each bi-text (109/116 aligned texts working actually[3]) the user has access to the alignment at the sentence and word level, and to the dictionary. "MultiSemCor+" (as defined by Lupu et al. (2005)) is a more recent extension that also contains the the Romanian SemCor (Section 2.1.3, Lupu et al., 2005). This new project represents a first test bed for multilingual semantic disambiguation experiments. We can browse the same aligned texts in Romanian and English on the MultiSemCor Browser. Currently the English-Romanian modality has only a subset of the Italian: 12/116 aligned texts.

#### 2.1.3 SemCor En-Ro corpus and RoSemCor

Even if the monolingual Romanian corpus is not so clearly available while the multilingual one is distributed open and free under MS Commons-BY-NC-ND [4]. En-Ro SemCor contains a total of 178,499 words for English and 175,603 words for Romanian (Lupu et al., 2005; Ion, 2007). The English SemCor texts have been translated into Romanian and the sentence and paragraph annotations have been observed. The sense transfer from English to Romanian follows closely the WSDTool procedure (a wordsense disambiguation algorithm described by Ion (2007)). From a total of 88,874 occurrences of content words in Romanian, 54.54% received sense annotation by the transfer procedure.

#### 2.1.4 Jsemcor

Japanese Sem-Cor (JSemCor: Bond et al., 2012) is a sense-tagged corpus for the Japanese Wordnet (Isahara et al., 2008), based on translation of the subset of English SemCor used in MultiSemCor (Section refsec:multisemcor) with senses projected across from English. In this case, of the 150,555 content words only 58,265 are sense tagged. Jsemcor is a SemCor corpus: the texts are aligned to the correspondent English SemCor texts both at the sentence and word level. The transfer process left 39% of the senses untagged because of the fundamental differences between Japanese and English. A major cause of lexical gaps is part-of-speech mismatches. The license is similar to the Princeton WordNet License, so the data is freely available.

### 2.2 Independent Corpora for other languages

Most projects sense-tag existing annotated corpora for their languages. This means that they can take advantage of the work that has gone into pre-processing them, and also be used with other annotations.

#### 2.2.1 BulSemCor

The Bulgarian Semantically Annotated Corpus (Koeva et al., 2010) is part of the Bulgarian Brown Corpus (balanced but not aligned to the English Brown Corpus, so BullSemCor is a NonSemCor corpus). It consists of 811 excerpts each containing 100+ words: the total size of the source corpus is 101,062 tokens.[5] Each lexical item (simple or compound word) which occurs in the particular context in BulSemCor is assigned manually the unique semantic or grammatical meaning from the Bulgarian wordnet. The result is a lemmatised POS and sense-annotated corpus of units of running text. Unlike most wordnet corpora, the annotation includes both open-class and closed-class words. Sense distinctions in the closed word classes have been drawn primarily from corpus evidence. The sense-annotated corpus consists of 99,480 lexical units annotated with the most appropriate synset from the Bulgarian wordnet (BulNet). The corpus excerpts are offered under MS NoRedistribution NonCommercial license [6] for free, it is also possible to query the corpus online. The restrictions on use and redistribution mean that corpus is not considered open source.

#### 2.2.2 Eusemcor and spsemcor

The University of the Basque Country and the Department of Software, Technical University of Catalonia have produced two browsing-online-only corpora: Eusemcor (Basque Semcor) and spsemcor (Spanish Semcor) (Agirre et al., 2006). Eusemcor was compiled with samples from a balanced corpus and a newspaper corpus. It comprises 300,000 words in total. Agirre et al. (2006) point out that as Basque is an agglutinative language, it has a higher lemma/word rate than English, so in parallel corpora it would allow to think that 300,000 words in Basque are comparable to 500.000 words in English. The process of tagging the new corpus was

[3] multisemcor.fbk.eu/frameset1.php

[4] http://meta-net.eu/meta-share/meta-share-licenses/META-SHARE=%20COMMONS_BYNCND%20v1.0.pdf

[5] dcl.bas.bg/en/corpora_en.html#SemC

[6] http://www.meta-net.eu/meta-share/meta-share-licenses/META-SHARE%20NonCommercial%20NoRedistribution%20NoDerivatives%20For-a-fee-v%201.0.pdf

used in this case mainly to extend the Basque WordNet adding the eventual missing needed senses. Spsemcor is a part of SenSem, a databank of Spanish which maps a corpus and a verbal database. The SenSem corpus consists of 25,000 sentences, 100 for each of the 250 most frequent verbs of Spanish (Davies, 2002). Sentences are tagged at both syntactic and semantic levels: verb sense, phrase and construction types, aspect, argument functions and semantic roles. In the Spsemcor part of SenSem the noun heads were tagged with the Spanish WordNet 1.6: 23,307 forms for 3,693 noun lemmas of the SenSem corpus have been semantically annotated (Climent et al., 2012). This corresponds to the 82.6% of the total amount of verbal arguments in the corpus. Both Eusemcor and Spsemcor are only available for online browsing.

#### 2.2.3 AnCora

AnCora (Martí et al., 2007) are two multilingual corpora of 500,000 words each: a Catalan corpus (AnCora-CAT) and a Spanish (AnCora-ESP) one, built in an incrementally way from the previous 3LB corpora.[7] In this way, 400,000 words were added to each corpus coming from different press sources (mainly newspapers). The AnCora corpora were annotated at different levels of linguistic description: the whole Catalan corpus is annotated with morphological, syntactic, and semantic information; as for Spanish, the morphological and syntactic levels are already completed, while the semantic annotation covers 40% of the corpus (200,000 words). The lexical semantic annotation consists in assigning each noun in the corpora its sense. This process was carried out manually and the senses repository is WordNet. Each noun was assigned either a WordNet sense or a label indicating a special circumstance.

#### 2.2.4 DutchSemCor

DutchSemCor is a sense-tagged corpus with senses and domain tags from the Cornetto lexical database (Vossen et al., 2011). In DutchSemCor about 282,503 tokens for 2,870 nouns, verbs and adjectives (11,982 senses) have been manually tagged by two annotators, resulting in 25 examples on average per sense (anyway more than 400,000 have been manually tagged by at least one annotator and millions have been automatically tagged). The examples mainly come from existing corpora collected in the projects CGN (9 millions words: Van Eerten, 2007), D-Coi, and SoNaR (500 millions words: Oostdijk, 2008), but also additional examples from the Dutch websites have been added. DutchSemCor is not available, but excerpts and statistics are freely downloadable.

[7] Read Civit and Martí (2004) for 3LB-ESP and Civit et al. (2004) for 3LB-CAT

#### 2.2.5 TüBa-D/Z Treebank

Henrich and Hinrichs (2013) have manually annotated the TüBa-D/Z Treebank[8] with GermaNet senses with the goal of providing a gold standard for word sense disambiguation. The underlying resource is a German newspaper corpus manually annotated at various levels of grammar. The sense inventory used for tagging word senses is taken from GermaNet. With the sense annotation for a selected set of 109 words (30 nouns and 79 verbs) occurring 17,910 times in the TüBa-D/Z, the treebank currently represents the largest manually sense-annotated corpus available for GermaNet. The corpus is not currently available but it will be made freely available in a future release at the TüBa-D/Z Sense Annotations webpage.[9]

#### 2.2.6 WebCaGe

WebCaGe is a web-harvested corpus annotated with GermaNet senses, the largest sense-annotated corpus available for German (Henrich et al., 2012). WebCaGe includes example sentences from the German Wiktionary (46,457 German words) and additional material collected by following the links to Wikipedia, the Gutenberg archive, and other web-based materials. Wiktionary (7,644 tagged word tokens) and Wikipedia (1,732) contribute by far the largest subsets of the total number of tagged word tokens (10,750) compared with the external webpages (589) and the Gutenberg texts (785). These tokens belong to 2,607 distinct polysemous words contained in GermaNet, among which there are 211 adjectives, 1,499 nouns, and 897 verbs. On average, these words have 2.9 senses in GermaNet (2.4 for adjectives, 2.6 for nouns, and 3.6 for verbs). WebCaGe is distributed under the Creative Commons Attribution-ShareAlike 3.0 Unported License (CC BY-SA 3.0) [10]

#### 2.2.7 ISST

ISST is the Italian Syntactic-Semantic Treebank (Montemagni et al., 2003) a multi-layered annotated corpus of Italian. ISST has a five-level structure covering orthographic, morpho-syntactic, syntactic and semantic levels of linguistic description. The fifth level deals with lexico-semantic annotation, which is carried out in terms of sense tagging of lexical heads (nouns, verbs and adjectives) augmented with other types of semantic information: ItalWordNet (Italian part of the EuroWordNet Project) is the reference lexical resource used for the sense tagging task. The ISST corpus consists of 305,547 word tokens (composing a balanced corpus for a total of 215,606 tokens and a specialized

[8] www.sfs.unituebingen.de/en/ascl/resources/corpora/tueba-dz.html

[9] http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/sense-annotated-tueba-dz.html

[10] http://creativecommons.org/licenses/by-sa/3.0/

corpus, amounting to 89,941 tokens, with texts belonging to the financial domain) of which 81,236 content words are sense annotated. ISST was made available for research purposes in 2010 (Dei Rossi et al., 2011).

#### 2.2.8 NTU-MC

The NTU-Multilingual Corpus is a corpus designed to be multilingual from the start. It contains parallel text in eight languages: English (eng), Mandarin Chinese (cmn), Japanese (cpn), Indonesian (ind), Korean (kor), Arabic (arb), Vietnamese (vie) and Thai (tha) (Tan and Bond, 2012). Text is in three genres: short stories, essays and tourism. All the text is translated from English. The text is being sense annotated (Open Multilingual Wordnet[11] senses) in Chinese, English, Japanese and Indonesian (tourist data only; Bond et al., 2013). Tagging is still underway, snapshots are available from `compling.hss.ntu.edu.sg/ntumc`. The sizes of the different subcorpora are given in Table 1. There is more data for Chinese and English, with less for Indonesian and Japanese.

#### 2.2.9 AQMAR Arabic SST

This is a 65,000-token corpus[12] of 28 Arabic Wikipedia articles (selected from the topical domains of history, sports, science, and technology) hand-annotated for nominal supersenses (40 coarse lexical semantic classes, 25 for nouns, 15 for verbs, originating in WordNet). It extends the Named Entity Corpus[13] and was developed by Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah Smith (Schneider et al., 2012) as part of the AQMAR project.[14] This dataset is released under the Creative Commons Attribution-ShareAlike 3.0 Unported license (CC BY-SA 3.0).

#### 2.2.10 Jos100k

The Jos100k corpus of Slovene contains 100,000 words of sampled paragraphs from the FidaPLUS corpus.[15] It is meant to serve as a reference annotated corpus of Slovene: its manually-validated annotations cover three level of linguistic description (morphosyntactic, syntactic and semantic). All the occurences of 100 most frequent nouns are annotated with their concept (synset id) from the Slovene WordNet sloWNet. The corpus is now at the version 2.0 and is freely available (CC BY-NC 3.0 [16]) for browsing and downloading at the project webpage: `nl.ijs.si/jos/jos100k-en.html`. An online browser for concordances is available here `nl.ijs.si/jos/cqp/` and a lot of documenting information is available as TEI corpus.[17]

[11] `compling.hss.ntu.edu.sg/omw`

[12] `www.ark.cs.cmu.edu/ArabicSST/`

[13] `www.ark.cs.cmu.edu/ArabicNER/`

[14] `www.ark.cs.cmu.edu/AQMAR/`

[15] `www.fidaplus.net/`

[16] `http://creativecommons.org/licenses/by-nc/3.0/deed.en`

[17] `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-teiCorpus.html`

#### 2.2.11 Hungarian word sense disambiguated corpus

The Hungarian WSD corpus (Vincze et al., 2008), contains 39 suitable word form samples selected (the most frequent words with more than one well-defined senses) for the purpose of word sense disambiguation. There are 300-500 samples for each word (so more or less 16,000 thousands samples). The Hungarian National Corpus and its Heti Világgazdaság (HVG) subcorpus provided the basis for corpus text selection and senses are from the Hungarian WordNet (HuWN) [18]. This corpus is a fine-grained lexical sample corpus.The corpus follows the SemEval XML format (not validable XML).

#### 2.2.12 KPWr Polish Corpus of Wroclaw University

The Polish Corpus of Wroclaw University (Broda et al., 2012) represents written and spoken Polish. All the documents are freely available under the Creative Commons Attribution 3.0 Unported Licence[19]. The texts are organized in 14 categories (blogs, science, stenographic recordings, dialogue, contemporary prose, past prose, law, long press articles, short press articles, popular science and textbooks, wikipedia, religion, official texts and technical texts). The annotations are on the level of chunks and selected predicate-argument relations, named entities, relations between named entities, anaphora relations and word senses (plwordnet[20] senses). The corpus contains totally 438,327 words with 9157 tagged (for selected lexems) and has been been developed by The WrocUT Language Technology Group G4.19, Artificial Intelligence Department at the Institute of Informatics, Wroclaw University of Technology.

### 2.3 Other English Corpora

As is common for language resources, there are more for English than for any other language.

#### 2.3.1 WordNet Gloss Corpus

In the Princeton WordNet Gloss Corpus Word,the definitions (or glosses) of WordNet's synsets are manually linked to the context-appropriate sense in WordNet. The corpus contains 1,621,129[21] tokens with 449,355 sense tagged (330,499 manually + 118,856 automatically) on 656,066 taggable words and globs (the tagged ones + 206,711 untagged). The wordnet definitions have been translated into many languages, including Albanian (Ruci, 2008), Japanese (Bond et al., 2010), Korean (Yoon et al., 2009) and Spanish (Fernández-Montraveta et al., 2008). Further, the glosses are useful for unsupervised sense disambiguation techniques such

[18] `http://www.inf.u-szeged.hu/rgai/HuWN`

[19] `http://creativecommons.org/licences/by/3.0/legalcode`

[20] `plwordnet.pwr.wroc.pl/wordnet`

[21] `wordnet.princeton.edu/glosstag.shtml`

as LESK (Lesk, 1986): and it has been shown for another resource that having the glosses disambiguated improves the accuracy of extended LESK (Baldwin et al., 2008).

#### 2.3.2 Groningen Meaning Bank

The Groningen Meaning Bank (GMB), is a free corpus of English (1,020,367 tokens) developed at the University of Groningen, comprises thousands of texts in raw and tokenised format, tags for part of speech, named entities and lexical categories (word senses from WordNet, among other things), and discourse representation structures compatible with first-order logic (Basile et al., 2012). The senses are mostly automatically annotated, though part of them are manually corrected through the GMB wiki-like interface: `gmb.let.rug.nl/explorer`. The current (development) version of the GMB is accessible via the GMB Explorer: everbody is explicitly invited to contribute to the GMB by providing corrections to existing linguistic annotations with the simplicity made possible by such a wiki-like environment. Anyone can register via the GMB Explorer and check, improve, or discuss linguistic annotations. Stable releases are made available periodically and are freely available from the downloads webpage. Data from the Wordrobe[22] platform is also used to correct word senses in the GMB, applying the very innovative crowdsourcing technique "Game with a Purpose" (GWAP): rewarding contributors with entertainment rather than money. The design and the first results of Wordrobe are presented in Venhuizen et al. (2013).

#### 2.3.3 MASC

MASC (Manually Annotated Sub-Corpus) is a part of the American National Corpus (Ide, 2012) with multiple layers of annotations in a common format that can be used either individually or together, and (unlike, for example, OntoNotes) to which others can add annotations. MASC currently contains nineteen genres of spoken and written language data in roughly equal amounts, covers a wide range of written genres, including emerging social media genres (tweets, blogs). The entire MASC is annotated for logical structure, token and sentence boundaries, part of speech and lemma, shallow parse (noun and verb chunks), named entities (person, location, organization, date), and Penn Treebank syntax. Portions of MASC are also annotated for additional phenomena, including 40,000 of full-text FrameNet frame element annotations and PropBank, TimeML, and opinion annotations over a roughly 50,000 subset of the data. MASC also includes sense-tags for 1,000 occurrences of each of 100 words chosen by the WordNet and FrameNet teams (100,000 annotated occurrences), described in (Ide, 2012). The sense-tagged data are distributed as a separate sentence corpus with links to the original documents in which they appear. Where MASC does not contain 1000 occurrences of a given word, additional sentences were drawn from the OANC. All annotations have either been manually produced or automatically produced and hand-validated. MASC is distributed without license or other restrictions.

#### 2.3.4 DSO Corpus of Sense-Tagged English

This sense tagged corpus was provided by Ng and Lee (1996) of the Defence Science Organisation (DSO) of Singapore and has been hand tagged by 12 undergraduates from the Linguistics Program of the National University of Singapore. It contains sense-tagged word occurrences for 121 nouns and 70 verbs which are among the most frequently occurring and ambiguous words in English. These sentences are taken from the Brown corpus and the Wall Street Journal corpus. About 192,800 word occurrences have been hand tagged with WordNet 1.5 senses. It is distributed on the Linguistic Data Consortium Catalogue[23] (LDC) under different licences for LDC Members (free for 1997 members) and Non-Members.

#### 2.3.5 OntoNotes

OntoNotes Release 5.0[24] is the final release of the OntoNotes project,[25] a collaborative effort between BBN Technologies, the University of Colorado, the University of Pennsylvania and the University of Southern Californias Information Sciences Institute. The goal of the project was to annotate a large corpus comprising various genres of text (news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows) in three languages (English, Chinese, and Arabic) with structural information (syntax and predicate argument structure) and shallow semantics (word sense linked to an ontology and coreference). OntoNotes Release 5.0 contains the content of earlier releases and adds source data from and/or additional annotations for, newswire (News), broadcast news (BN), broadcast conversation (BC), telephone conversation (Tele) and web data (Web) in English and Chinese and newswire data in Arabic. Also contained is English pivot text (Old Testament and New Testament text). This cumulative publication consists of 2.9 million words. Its semantic representation includes word sense disambiguation for nouns and verbs. The sense annotation is done on coarse grained clusters of wordnet senses (OntoNotes Sense Groups) for 1.5 million words of English.

#### 2.3.6 SemLink

SemLink is a project whose aim is to link together different lexical resources via set of mappings. These mappings could make it possible to combine the different information provided by these different lexical

[22] `gmb.let.rug.nl/wordrobe.php`

[23] `catalog.ldc.upenn.edu/LDC97T12`

[24] `catalog.ldc.upenn.edu/LDC2013T19`

[25] `www.bbn.com/ontonotes/`

resources for tasks such as inferencing. Currently SemLink contains mappings between PropBank,[26] VerbNet,[27] FrameNet[28] and WordNet[29](which is again represented by the OntoNotes Sense Groups). The content of all four of these resources can be browsed on-line using the Unified Verb Index.[30] The SemLink corpus is the WSJ portion of the Penn TreeBank, currently at Version 1.2.2c with approximately 78,000 tokens. The corpus is freely downloadable and browsable on the SemLink project webpage.[31]

### 2.4 Senseval and SemEval tasks and lexical samples

SemEval (Semantic Evaluation) is an ongoing series of evaluations of computational semantic analysis systems. The first three evaluations, Senseval-1 through Senseval-3, were focused on word sense disambiguation, then Senseval evolved from the Senseval word sense evaluation series to the new SemEval series. In fact during the fourth workshop, SemEval-2007 (SemEval-1), the nature of the tasks evolved to include semantic analysis tasks outside of word sense disambiguation. Each of these evaluations provided some lexical samples or little corpora. Here we list the most recent and relevant.

#### 2.4.1 Senseval 1-3

The first SENSEVAL took place in 1998, for English, French and Italian, culminating in a workshop. Senseval 1[32] provided a corpus containing 12,000+ instances of 35 words, and a practice run corpus distributed prior to Senseval 1, containing 20,000+ instances of 38 words. In 2001 Senseval 2 provided a corpus containing 12,000+ instances of 73 words. For the "English all-words task" at the Senseval-3, Snyder and Palmer (2005) prepared a sense-tagged corpus: 5,000 words from two Wall Street Journal articles (editorial domain the first, news story the second one) and one excerpt from the Brown Corpus (fiction). All verbs, nouns and adjectives have been double annotated with WordNet 1.7.1 senses, and then adjudicated and corrected by a third person. The total tagged words are 2,212 (given that some of these are multiwords the total number of tags is 2,081). All the data (ill-formed XML) produced for Senseval are freely available at the Senseval web page, but are also available at the Pedersen's webpage [33] in a partially corrected but still ill-formed XML version.

#### 2.4.2 Line, Hard, Serve and Interest Corpora

Pedersen has also collected and converted to the Senseval 2 format the corpora for ***line***, ***hard*** and ***serve***, each with 4,000+ noun instances, tagged with 6, 3 and 4 wordnet senses repectively Leacock et al. (1993), along with the ***interest*** corpus (2,369 instances from the ACL/DCI Treebank tagged with 6 LDOCE senses described by Bruce and Wiebe (1994)). All these resources are freely available at the Ted Pedersen's webpage[34].

#### 2.4.3 SemEval07–13

Many other resources are available at the SemEval2007[35], SemEval2010[36], SemEval2012[37] and SemEval2013[38] websites. In particular we have to mention Semeval-2013 Task 12 (all nouns tagged with WordNet 3.0 senses) and SemEval-2013 Task 13. The Task 12 test set consisted of 13 articles (Navigli et al., 2013) obtained from the datasets available from the 2010, 2011 and 2012 editions of the workshop on Statistical Machine Translation (WSMT). The articles cover different domains, ranging from sports to financial news. The same article was available in 4 different languages (English, French, German and Spanish). In order to cover Italian, an Italian native speaker manually translated each article from English into Italian, with the support of an English mother tongue advisor. In Table 1 we show for each language the number of words of running text, together with the number of multiword expressions and named entities annotated, from the 13 articles. The Task 13 (Jurgens and Klapaftis, 2013) has a lexical sample corpus for 20 nouns, 20 verbs, and 10 adjectives, tagged with WordNet 3.1 senses. In the dataset there are 4664 instances (on 141k tokens) and will soon be available on its task website[39]. Task 13's dataset (Jurgens and Klapaftis, 2013) covers multiple genres of text (spoken, newswire, fiction, etc.) and has annotations when multiple senses apply, with around 11% annotated with at least two senses that are weighted by applicability.

## 3 Discussion

Currently, there is no widely adopted format for wordnet annotated corpora (even if the ISO TC37/SC4 group[40] is working on the principles of semantic annotation[41]): every institution uses its own format, and very little sharing of tools to manipulate the data. This is despite much work on corpus standards. With the

[26] verbs.colorado.edu/~mpalmer/projects/ace.html

[27] verbs.colorado.edu/~mpalmer/projects/verbnet.html

[28] framenet.icsi.berkeley.edu/fndrupal/

[29] wordnet.princeton.edu/

[30] verbs.colorado.edu/verb-index/

[31] verbs.colorado.edu/semlink/

[32] www.senseval.org/

[33] www.d.umn.edu/~tpederse/data.html

[34] www.d.umn.edu/~tpederse/data.html

[35] www.senseval.org/

[36] semeval2.fbk.eu/semeval2.php?location=data

[37] www.cs.york.ac.uk/semeval-2012/

[38] www.cs.york.ac.uk/semeval-2013/

[39] www.aclweb.org/anthology/S/S13/S13-2049.pdf

[40] www.tc37sc4.org/index.php

[41] www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=60581

exception of the MultiWordNet, the corpora are not linked with the wordnets in an online interface. For those languages with sense tagged corpora, there are generally between 10–100 thousand tagged entries: far fewer than the number of senses in the wordnets. This means that most wordnet entries have no example in the corpus. Kilgariff and Rosenzweig (2000) argued that tagging all words was not useful from the lexicographers point of view: it is better to have 50-100 examples for each word, than 1 or 2 for many. However, for research into lexical semantics and the distribution of words, as well as the use of semantic classes as backoff in other processing, it is necessary to tag all words. This is the most common form of annotation. Most projects point out that the much of the time spent in annotation is in fact in adding new word senses — this is still a very hard problem.

English has the most sense tagged data, followed by Dutch, then Italian, Japanese and Romanian (assuming that much of the Bulgarian is closed class words). The last three are all tagging through projection — this is an efficient way to bootstrap sense annotation.

There are two projects that have created multilingual corpora. The first is the MultiSemCor project, which grew out of the MultiWordNet. Construction of multiple wordnets and corpora went hand in hand. They inspired a similar approach for Japanese. Their MultiSemCor Browser (Ranieri et al., 2004) is probably the best and most useful tool for researchers interested in studying multilingual information. Even so, there is still much to do. There are only two non-English corpora currently available and the browser works only with English-Italian/Romanian: there are no links between Italian and Romanian.

Building a new translated semcor is difficult for at least three reasons. The first problem is that the wordnet annotated corpora don't update their sense tagging system (based on a precise wordnet version) when the English WordNet and SemCor do. If your wordnet is linked to a different version, in order to combine them into a single multilingual structure, we have to map to a common version.

The second problem is the variety of formats used. So sometimes even if a corpus is legally available, there could be still a technical hurdle before it becomes easily accessible. Conversion to a common format is the obvious solution. Finally, translating SemCor is in itself expensive, even though it may be worth it due to the richness of the existing annotation that can be projected across.

The second multi-lingual project is the NTU Multilingual Corpus. Instead of translating an existing sense tagged corpus, they chose to choose texts already freely available in multiple languages, and use the translations to guide the annotation. This was more expensive to annotate at first, but has the potential to cheaply expand to more languages: projecting from the existing annotations.

One possible explanation for the lack of coordination in tools and formats is that many of the large corpora are not open-source (Dutch, DSO, Romanian, Spanish, Basque, WebCaGe, ISST). It is therefore not legally possible for people to reformat and redistribute the corpora. In contrast, the open English corpora have been mapped to the latest version of Wordnet and the same format and made available.[42] As more corpora are released under open licenses, we expect this state to improve.

## 4 Future Work

We would like to further the usefulness of the multilingual corpora in several ways. The first is to align the English, Italian, Romanian and Japanese translations of SemCor. We will then use English as a pivot to link Italian, Romanian and Japanese. When all four languages are aligned, we can use the translations to disambiguate and check the senses, as well as trying to make the projection more robust. The second is to do this with the NTU-multilingual corpus: make it compatible with MultiSemCor, align through English and refine. This will make it easier to add other languages: the Sherlock Holmes short stories and the Cathedral and the Bazaar have many translations. The third is to do this with the Wordnet Gloss Corpus: linking definitions in other languages to make a multilingual gloss corpus. It would also be interesting to use definitions from other sources (such as Wiktionary) to make an aligned sense-tagged paraphrase corpus. Finally (or in parallel) we would like to make these corpora all searchable, and linked to the Wordnet Grid (Pease et al., 2008; Bond and Foster, 2013).

## 5 Conclusions

All these observations about the compatibility troubles in the construction process of multilingual wordnet annotated corpora point at a clear fact: the more we standardize our data formats, and the more we open and share freely our resources and tools the easier and the faster will be the development of new resources all over the world.

**Acknowledgments**

This research was supported in part by the Erasmus Mundus Action 2 program MULTI of the European Union, grant agreement number 2009-5259-5. We would like to thank Anja Weisscher and Piek Vossen for their help in adding the information to the Global Wordnet Association page. We would also like to thank Shan Wang, Verena Henrich, Marie Hinrichs, Valerio Basile, Behrang M., Christiane D. Fellbaum, Ng Hwee Tou, David Jurgens, Mathieu Lafourcade, Orin Hargraves, Tomaz Erjavec, Vincze Veronika and Marcin Oleksy for their help and information.

[42] You can find SemCor, Senseval 2 and 3 here, `www.cse.unt.edu/~rada/downloads.html#semcor`

## References


Eneko Agirre, Izaskun Aldezabal, Jone Etxeberria, Eli Izagirre, Karmele Mendizabal, Eli Pociello, and Mikel Quintian. 2006. Improving the Basque wordnet by corpus annotation. In *In Proceedings of Third International WordNet Conference*, pages 287–290. Jeju Island, Korea.

Eneko Agirre, Timothy Baldwin, and David Martinez. 2008. Improving parsing and pp attachment performance with sense information. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL HLT 2008)*, pages 317–325. Columbus, USA.

Timothy Baldwin, Su Nam Kim, Francis Bond, Sanae Fujita, David Martinez, and Takaaki Tanaka. 2008. MRD-based word sense disambiguation: Further extending Lesk. In *International Joint Conference on Natural Language Processing 2008*, pages 775–780. Hyderabad, India.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In *LREC*, volume 12, pages 3196–3200.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the multisemcor corpus. *Natural Language Engineering*, 11(3):247–261.

Daniel M. Bikel. 2000. A statistical model for parsing and word-sense disambiguation. In *Student Research Workshop at ACL 2000*, pages 1–7. Hong Kong.

Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. Japanese semcor: A sense-tagged corpus of japanese. In *Proceedings of the 6th International Conference of the Global WordNet Association (GWC)*.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51th Annual Meeting of the Association for Computational Linguistics and the Human Language Technologies*, pages 1352–1362. Sofia. URL `http://aclweb.org/anthology/P13-1133`.

Francis Bond, Hitoshi Isahara, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2010. Japanese WordNet 1.0. In *16th Annual Meeting of the Association for Natural Language Processing*, pages A5–3. Tokyo.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue. 64–71.

Francis Bond, Shan Wang, Eshley Huini Gao, Hazel Shuwen Mok, and Jeanette Yiwen Tan. 2013. Developing parallel sense-tagged corpora with wordnets. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, page 149–158. Association for Computational Linguistics, Sofia, Bulgaria.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of LREC'12*. ELRA, Istanbul, Turkey.

F. Rebecca Bruce and M. Janyce Wiebe. 1994. Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 139–146.

Monsterrat Civit, Núria Bufı, and Pilar Valverde. 2004. Building cat3lb: a treebank for catalan. In *Proceedings of the SALTMIL Workshop at LREC 2004*, pages 48–51.

Montserrat Civit and Ma Antònia Martí. 2004. Building cast3lb: A spanish treebank. *Research on Language and Computation*, 2(4):549–574.

Salvador Climent, Marta Coll-Florit, Marina Lloberes, and German Rigau. 2012. Semantic hand tagging of the SenSem corpus using Spanish wordnet senses. In *GWC 2012 6th International Global Wordnet Conference*, page 72.

Mark Davies. 2002. Un corpus anotado de 100.000. 000 palabras del español histórico y moderno. *Procesamiento del lenguaje natural*, 29:21–27.

Stefano Dei Rossi, Giulia Di Pietro, and Maria Simi. 2011. Evalita 2011: Description and results of the supersense tagging task. *Evalita 2011*.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Ana Fernández-Montraveta, Gloria Vázquez, and Christiane Fellbaum. 2008. The spanish version of wordnet 3.0. *Text Resources and Lexical Knowledge. Mouton de Gruyter*, pages 175–182.

Dan Flickinger. 2011. Accuracy vs. robustness in grammar engineering. In E. M. Bender and J. E. Arnold, editors, *Language from a Cognitive Perspective: Grammar, Usage, and Processing*, pages 31–50. CSLI Publications.

W. Nelson Francis and Henry Kucera. 1979. *BROWN CORPUS MANUAL*. Brown University, Rhode Island, third edition. (`http://khnt.aksis.uib.no/icame/manuals/brown/`).

Verena Henrich and Erhard Hinrichs. 2013. Extending the tüba-d/z treebank with germanet sense annotation. In Iryna Gurevych, Chris Biemann, and Torsten Zesch, editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 89–96. Springer Berlin Heidelberg. URL `http://dx.doi.org/10.1007/978-3-642-40722-2_9`.

Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. 2012. Webcage: a web-harvested corpus annotated with germanet senses. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 387–396. Association for Computational Linguistics, EACL, Avignon, France.

Nancy Ide. 2012. Multimasc: An open linguistic infrastructure for language research. In *Proceedings of the Fifth Workshop on Building and Using Comparable Corpora*. Istanbul.

Radu Ion. 2007. *Metode de dezambiguizare semantica automata. Aplicat ii pentru limbile englezas i romana*. Ph.D. thesis, ACADEMIA ROMANA, Institutul de Cercet ari pentru Inteligenta Artificiala, Bucarest.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.

Toru Ishida. 2006. Language grid: An infrastructure for intercultural collaboration. In *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)*, pages 96–100. URL `http://langrid.nict.go.jp/file/langrid20060211.pdf`, (keynote address).

David Jurgens and Ioannis Klapaftis. 2013. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *Proceedings of the 7th International Workshop on Semantic Evaluation*.

Adam Kilgariff and Joseph Rosenzweig. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities*, 34(1–2):15–48. Special Issue on SENSEVAL.

Svetla Koeva, Svetlozara Leseva, Ekaterina Tarpomanova, Borislav Rizov, Tsvetana Dimitrova, and Hristina Kukova. 2010. Bulgarian sense annotated corpus - results and achievements. In M. Tadić, M. Dimitrova-Vulchanova, and S. Koeva, editors, *Proceedings of the 7th International Conference of Formal Approaches to South Slavic and Balkan Languages*, pages 41–48. FASSBL-7, Dubrovnik, Croatia.

Shari Landes, Claudia Leacock, and Christiane Fellbaum. 1998. Building semantic concordances. In Fellbaum (1998), chapter 8, pages 199–216.

Helen Langone, Benjamin R. Haskell, and George A. Miller. 2004. Annotating wordnet. In *Workshop On Frontiers In Corpus Annotation*, pages 63–69. ACL, Boston.

C. Leacock, G. Towell, and E. Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 260–265.

Michael Lesk. 1986. Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference*, pages 24–26. ACM, New York.

Monica Lupu, Diana Trandabat, and Maria Husarciuc. 2005. A Romanian semcor aligned to the English and Italian multisemcor. In *Proceedings 1st ROMANCE FrameNet Workshop at EUROLAN 2005 Summer School*, pages 20–27. EUROLAN, Cluj-Napoca, Romania.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational linguistics*, 19(2):313–330.

Maria Antònia Martí, Mariona Taulé, Manu Bertran, and Lluís Màrquez. 2007. Ancora: Multilingual and multilevel annotated corpora. *MS, Universitat de Barcelona*.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 240–243. ARPA.

Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, MariaTeresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, and Rodolfo Delmonte. 2003. Building the Italian syntactic-semantic treebank. In Anne Abeillé, editor, *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 189–210. Springer Netherlands.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (* SEM 2013), Atlanta, Georgia*, pages 14–15.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 40–47.

Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christoper D. Manning. 2004. LinGO redwoods: A rich and dynamic treebank for HPSG. *Research on Language and Computation*, 2(4):575–596.

NHJ Oostdijk. 2008. Sonar: Stevin nederlandstalig referentiecorpus.

Adam Pease, Christine Fellbaum, and Piek Vossen. 2008. Building the global wordnet grid. In *Proceedings of the CIL-18 Workshop on Linguistic Studies of Ontology*. Seoul. URL `http://www.adampease.org/professional/Grid2008.pdf`.

Christopher Mark Powell. 2005. *From E-Language to I-Language: Foundations of a Pre-Processor for the Construction Integration Model*. Ph.D. thesis, Oxford Brookes University.

Marcello Ranieri, Emanuele Pianta, and Luisa Bentivogli. 2004. Browsing multilingual information with the multisemcor web interface. In *Proceedings of the LREC 2004 Satellite Workshop on The Amazing Utility of Parallel and Comparable Corpora*, pages 38–41. LREC.

Ervin Ruci. 2008. On the current state of Albanet and related applications. Technical report, University of Vlora. (`http://fjalnet.com/technicalreportalbanet.pdf`).

Geoffrey Sampson. 1995. In *English for the Computer: The SUSANNE Corpus and Analytic Scheme*, pages +499 pp. Oxford: Clarendon Press, University of Sussex.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A Smith. 2012. Coarse lexical semantic annotation with supersenses: an arabic case study. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 253–258. Association for Computational Linguistics.

Benjamin Snyder and Martha Palmer. 2005. The english all-words task. In *Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, pages 41–43.

Liling Tan and Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing*, 22(4):161–174.

Laura Van Eerten. 2007. Over het corpus gesproken nederlands. *Nederlandse Taalkunde*, 12(3):194–215.

Noortje J Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. 2013. Gamification for word sense labeling. In *Proc. 10th International Conference on Computational Semantics (IWCS-2013)*, pages 397–403.

Veronika Vincze, György Szarvas, Attila Almási, Dóra Szauter, Róbert Ormándi, Richárd Farkas, Csaba Hatvani, and János Csirik. 2008. Hungarian word-sense disambiguated corpus. In *LREC*.

Piek Vossen, Attila Görög, Fons Laan, Maarten van Gompel, Rubén Izquierdo, and Antal van den Bosch. 2011. Dutchsemcor: building a semantically annotated corpus for Dutch. *Proceedings of eLex*, pages 286–296.

Aesun Yoon, Soonhee Hwang, Eunroung Lee, and Hyuk-Chul Kwon. 2009. Construction of Korean wordnet KorLex 1.5. *Journal of KIISE: Software and Applications*, 36(1):92–108.