Computational Linguistics in the Netherlands Journal 3 (2013) 49-62 Submitted 05/2013; Published 12/2013

DutchSemCor: Aiming at the ideal sense-tagged corpus

Piek Vossen PIEK.VOSSEN@VU.NL
Rubén Izquierdo RUBEN.IZQUIERDOBEVIA@VU.NL
Attila Görög A.GOROG@VU.NL

Computational Lexicology & Terminology Lab
Vrije Universiteit Amsterdam
The Netherlands

Abstract

The most-frequent-sense and the predominant domain sense play an important role in the debate on Word Sense Disambiguation (WSD). This discussion is, however, biased by the way sense-tagged corpora are built. In this paper, we argue that current sense-tagged corpora neglect rare senses and contexts and, as a result, do not represent a good corpus for training and testing word-sense-disambiguation. We defined three quality criteria for sense-tagged corpora and a methodology to satisfy these criteria with minimal effort. Following this method, we built a Dutch sense-tagged corpus that tried to meet these criteria. The corpus was evaluated by deriving word-sense-disambiguation systems and testing these on different subsets of the corpus in various ways. The performance of our systems and the quality of the derived data are equal to state-of-the-art English systems and corpora. Finally, we used the systems to annotate a chunk of the Dutch SoNaR-corpus and create a subcorpus of over 47 million sense-tagged tokens spread over a large variety of genres, domains and usages of Dutch. The results of the project can be downloaded freely from the project website.

1. Introduction

Word Sense Disambiguation (WSD) research in the last decade demonstrated a number of important insights (Agirre and Edmonds 2006): 1. evaluation results are strongly dependent on the corpus and the lexicons used, 2. the most-frequent-sense derived from SemCor (Miller et al. 1993) is a strong baseline that is not easy to beat in evaluations like SensEval or SemEval and 3. predominant senses in specific domains give the best WSD results by far (McCarthy et al. 2007). From these observations, one may conclude that we need to collect large sets of (sense-tagged) domain- and probably genre-specific corpora to determine predominant senses. Obtaining sufficient data without ignoring rare or low-frequent senses, however, requires an enormous effort. Manually tagged data is still very sparse and evaluation results vary from task to task, hence we still do not know where we stand in the area of WSD.

This raises the question: what should the ideal sense-tagged corpus for WSD look like, to enable detection of any sense in any type of corpus? Existing sense-tagged corpora have different design properties that make them good corpora in some aspects but not in others. In this paper, we will define quality criteria for sense-tagged corpora and will describe a novel method for building a large-scale sense-tagged Dutch corpus that meets these criteria with as little manual effort as possible. We argue that an ideal sense-tagged corpus should be balanced for the different senses (same number of annotations for each sense), for the different contexts (same number of annotations for each context) and should provide information on sense-frequencies, preferably across a wide range of domains and genres. These three characteristics are usually contradictory and a compromise solution will be addressed.

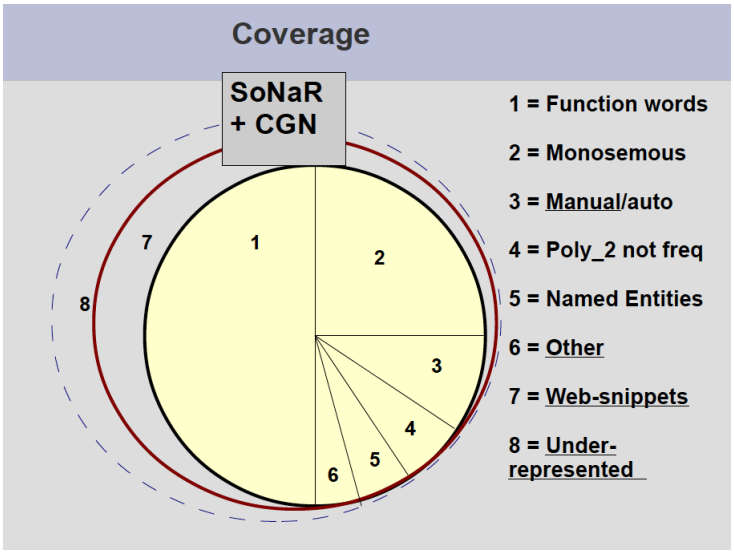

Figure 1: Coverage of senses.

In the DutchSemCor project we tried to meet these three criteria by using large corpora that cover a wide range of language-use, including spoken and written language, Flemish and Dutch standard language and dialects, and numerous genres and domains. Furthermore, we tagged these corpora through a mixture of manual and automatic annotations and selections of word tokens. We first aimed at a corpus that represents the meanings of an existing lexicon including sufficient examples for rare senses. Secondly, we extended this corpus to acquire a wider representation of contexts when needed and, finally, in order to acquire sense-distributions, the full corpus was annotated automatically applying three WSD systems. In Figure 1 the coverage of the annotated tokens can be seen.

The resulting annotations (both manual and automatic) were tested for all three criteria. As a side result, we obtained three WSD systems for Dutch that can be freely used for research and that perform at state-of-the-art level of English WSD systems.

The paper is structured as follows. In Section 2, we describe related work and different types of sense-tagged corpora that are commonly used. After a discussion of the advantages and disadvantages of each type of corpus, we define the main criteria that a sense-tagged corpus should meet. In Section 3, we outline our overall approach. In Section 4, a short overview of the resources (tools and corpora) is given. We describe the different phases of the annotation process including their evaluation in the subsequent sections: 5, 6, 7. In Section 8, we discuss the overall results.

2. Related work

Roughly speaking, there are two methods to annotate a corpus with senses: **sequential tagging** and **targeted tagging**. In the case of **sequential tagging**, annotators read a text word by word while annotating each occurrence. In the case of **targeted tagging**, the annotators will get a list (usually a KWIC index) of sentences for a single word and they annotate all the occurrences of the word. In the former case, annotators read each context only once but they need to reconsider the possible meaning of a word over and over again, each time they come across it. In the latter case, the annotators can tag all the occurrences of a word in one task and even apply contrastive analysis when considering all the contexts. The drawback is that they may have to read the same context

again when another word of the same context is annotated. The two approaches usually produce different annotation results for the same text and usually **targeted tagging** is more systematic and faster.

In addition to the annotation method, we can also distinguish sense-tagged corpora by their textual coverage. **Sequential tagging** usually results in an **all-words corpus** that contains annotations for all content words in texts. **Targeted tagging** usually results in a **lexical sample corpus**, a selection of target word occurrences with different contexts annotated with senses. The most famous example of an **all-words corpus** is SemCor (Miller et al. 1993), which was created through **sequential tagging** of parts of the Brown corpus (186 texts have all-words annotation, while in 166 texts only the verbs are annotated). An example of a **lexical-sample corpus** is the so-called line-hard-serve corpus (Mooney 1996)[1], which contains 4,000 instances of the noun *line* (six meanings), 4,000 instances of the verb *serve* (four meanings), and 4,000 instances of the adjective *hard* (three meanings).

Another **lexical-sample corpus** is DSO(Ng and Lee 1997) which has annotations only for the most frequent and ambiguous nouns (121) and verbs (70) in parts of the Brown corpus and a selection of Wall Street Journal articles, but is comparable in size to SemCor. For evaluation purposes, many other small **all-words** and **lexical-sample corpora** have been produced (cf. Senseval and SemEval competitions).

Lexical-sample and **all-words corpora** can often differ in the range and selection of their texts. Usually, **all-words corpora** cover a small number of texts, limited genres and domains and, as a result, a small number of senses, while **lexical sample corpora** usually represent a large number of different contexts and meanings of the target word. SemCor and DSO partly inherit the balanced nature of the Brown corpus. The corpora used in the Senseval evaluations: BNC, Wall Street Journal, Penn Treebank, part of Brown, show a variety of text types but do not provide systematic coverage neither of senses nor of different text types. Not surprisingly, the evaluation results of the Senseval competitions vary with the variation of corpora[2]. The lexical sample results vary from 64% to 77% and the all-words results vary from 45% to 69% (Agirre and Edmonds 2006). Interestingly, the inter-annotator-agreements (IAA) also vary a lot across the different tasks: 67% to 86% for the lexical sample tasks and 62% to 75% for the all-words task, as reported by (Agirre and Edmonds 2006). In all the competitions, the most-frequent-sense (MFS) in SemCor turned out to be a strong baseline (used as a fallback by many systems) that scores only a few points below the best systems (Agirre and Edmonds 2006).

These results raise a number of questions on how to annotate corpora with senses and how to develop WSD systems. Are the corpora for training and testing diverse enough in terms of contexts since they show so much variation in results? If MFS defines the ceiling for most systems, does this imply that we are neglecting low-frequent senses? Very often, annotators choose for representing the corpus used for the annotation rather than representing the sense repository used for annotating. Consequently, low frequent senses are not well represented in the training data. Besides, systems (and often also the evaluations) are too much skewed towards the most frequent senses. Depending on the evaluation set, a corpus that is not balanced for the different senses could give totally different results.

3. Our overall approach

We believe that sense-tagged corpora should be designed more carefully to provide answers to the above questions. We suggest three criteria for a sense-tagged corpus:

1. See also the *interest* (Bruce and Wiebe 1994) corpus.
2. Only Senseval-1 used a different lexical database. Senseval2&3 used WordNet1.7 and subsequent competitions used other versions of WordNet (Fellbaum 1998).

1. balanced-sense corpus: provide tokens and contexts for words that clearly illustrate the meaning of a word and provide equal numbers of examples for each meaning;

2. balanced-context corpus: provide tokens and contexts that represent the different usages of words in a representative corpus;

3. sense-probability corpus: provide a representative sample of the true frequency of a word meaning in a representative corpus.

To get a balanced-sense (1) and balanced-context (2) corpus, annotators need to build a lexical sample corpus by selecting or searching examples that fit the given senses best, where they can ignore unclear and problematic tokens of a word and avoid annotating the same contexts twice. To get a sense-probability corpus, a representative sample of language use from different styles, genres and domains needs to be annotated. The annotators have to assign senses to all the tokens selected by the sampler and they cannot discard tokens.

Obviously, the larger an annotated corpus the better. The question is how to build a corpus that tries to meet the above criteria using as little manual effort as possible. We propose a mixture of manual and automatic annotations:

1. Manually create a balanced-sense corpus (criterion 1). This corpus has an equal number of corpus examples for each sense, also for rare senses, and as-much-as-possible representing the variety of contexts rather than predominantly selecting examples with the same context.

2. Use this lexical sample corpus to train a WSD system that automatically annotates the remainder of a very large and diverse corpus. This corpus represents a large variety of contexts (criterion 2), while the WSD does not suffer from over-fitting for the MFS or for contexts and properties of the training corpus. Likewise, the system can detect rare senses equally well as frequent senses.

3. We use the complete set of annotations (manual and automatic) to obtain information on the sense-distributions (criterion 3) and to develop an MFS approach.

4. We evaluate a random sample of the tagged corpus to evaluate the automatic annotation and we test the WSD and the MFS on an all-words evaluation set. This will tell us how well the automatic annotation through the WSD system can handle the different contexts and how well it reflects the sense distributions.

Below, we will describe how we implemented this approach in the DutchSemCor project and what the results are. In the next section, we will first describe the resources we used.

4. Resources

We used the Cornetto database (Vossen et al. 2007) as the sense repository for the annotation. Cornetto combines a Dutch WordNet database with a traditional lexical-unit database that has detailed information on lexical units (synonyms in the Dutch WordNet). For the annotation, we made a selection of the 2,870 most polysemous and frequent content words in the database. The words together represent 11,982 word meanings with an average polysemy of around 3 senses per word. Figure 2 shows the information available in the annotated examples.

This is an example of the annotation of a sentence, that corresponds with the English sentence: *Kasparov chose another method to harass the horse on f5.* The semantic information is annotated at different levels. At the sentence level, these domains are automatically assigned to the sentence (with the corresponding confidence value): biology (biol), media, politics (pol), art (kunst), military (mil), linguistics (taal), transport (trans), commerce (handel), sports (sp) and agriculture (landb).

Figure 2: Annotated example sentence with domain, entity and sense information.

The token "Kasparov" is detected as a Named Entity and it is linked with the wikipedia page for Garri Kasparov. Finally the token "paard" (horse) is automatically disambiguated by three systems explained later in this section, and the guessed meaning by each system is assigned to the token. The sense with label *r_n-27277* stands for the meaning of horse as a vaulting horse in gymnastics, while the sense *r_n-27278* represents horse as a chess piece.

As our primary corpus, we used the SoNaR corpus (Oostdijk et al. 2013), which contains circa 500 million tokens of written Dutch and covers a wide range of different genres and topics (34 different categories including discussion lists, subtitles, books, legal texts, sms, chats, autocues, etc). SoNaR is fully tokenised, part-of-speech tagged, and lemmatised. Another corpus used was CGN (Corpus Gesproken Nederlands) which contains about nine million words of transcribed spontaneous Dutch adult speech. SoNaR is a very large corpus; however, it appeared not big enough to offer sufficient examples for a number of possibly rare senses (even if lexicographers agreed that these senses did exist). In Table 1 we can see the senses that required a greater amount of examples from the Web because they were not found in SoNaR, or were not clear enough.

PoS	Lemma	Sense	Freq.	English desc.
nouns	bult	r_n-8814	bump: a lump on the body caused by a blow	78
	staart	r_n-35354	tail: end part of a machine	76
	boer	r_n-7838	eructation	65
verbs	omslaan	c_546629	flip: to turn a page	102
	afronden	r_v-68	round: to make a round number	60
	houden	c_545697	keep: to remain in the standard referred to	59
adjs	betrokken	r_a-9258	cloudy	37
	schraal	r_a-14849	chapped skin or hands ...	33
	rationeel	r_a-14748	rational numbers	33

Table 1: Senses with the largest number of instances from the Web.

As can be seen, most of the these meanings are quite specific or are used in contexts that probably are not contained in SoNaR, or at least not very good examples for these senses. For this reason, we developed a tool in order to search additional examples on the Web through the WebCorp platform[3]. The annotators could make a selection of Internet examples and add these to the corpus. The web-snippets were then automatically tokenised, part-of-speech tagged and lemmatised. The final DutchSemCor corpus is, thus, a superset of SoNaR, CGN, and the manually-selected web-snippets.

During the project, we developed three Word-sense-disambiguation (WSD) systems, all three based on Machine Learning. The first one, called **DSC-TiMBL**, is a supervised Machine Learning system based on TiMBL (Daelemans et al. 2007). It implements a K-nearest neighbor algorithm (Aha et al. 1991). TiMBL has been widely used in NLP tasks. In the project, we used three different types of features. From the local context, we selected the word forms, lemmas and part-of-speech tags. The global context was modelled through bag-of-words contained in the same sentence as the target word. Finally, the system made use of information on SoNaR text type and of the token identifier to which the example belonged. Some filtering for the bag-of-words was performed in order to ensure the quality of the word predictors following the approach in (Ng and Lee 1996), which basically takes into account the frequency of a word in the context of one sense compared to the global frequency of that word, and discards very low frequency words.

The second system (**DSC-SVM**) uses a supervised Machine Learning approach based on Support Vector Machines, which belongs to the family of linear separators (Cortes and Vapnik 1995). This technique was extensively used in automatic classification tasks applying WSD systems and showed excellent performance in very high dimensional and sparse feature spaces, which is typically the case for WSD. In the project, we used the library SVMLight[4]. In this case the features were a bag-of-words around the target words. We also carried out a filtering process similar to the one mentioned on the previous paragraph.

The third system (**DSC-UKB**) was an unsupervised Machine Learning system based on the UKB algorithm (Agirre and Soroa 2009). This algorithm implements a so-called Personalized Page Rank algorithm similar to the one used by Google. It considers WordNet as a graph where each synset is a node in the graph and the relation between the synsets are seen as edges between the nodes. Disambiguation is performed through the ranking of the candidate nodes following the Personalized Page Rank algorithm. We used different sets of relations to build the graph: relations of the Dutch WordNet, English WordNet, equivalence relations from Dutch synsets to English synsets, WordNet Domain relations and co-occurrence relations extracted from the manual annotations of our corpus (i.e. relations between monosemous words and annotated polysemous examples)[5].

5. Building a balanced-sense corpus

To create a balanced-sense corpus, a team of annotators (trained student assistants) used an annotation tool developed within the project (SAT for Semantinc Annotation Tool) that loads data on the word meanings from the Cornetto lexical database and examples from the corpora mentioned in Section 4. A screenshot of the SAT tool can be seen in Figure 3.

The annotators could use various search strategies to find examples matching the selected meanings. Annotators needed to reach a high (80% or higher) Inter Annotator Agreement (IAA) and were instructed to select 25 good and clear examples per sense.

3. `http://www.webcorp.org.uk/live`
4. `http://svmlight.joachims.org`
5. 1.8 million relations were used in total: 1 million derived from Cornetto and WordNet and 800,000 derived from the manually-tagged data.

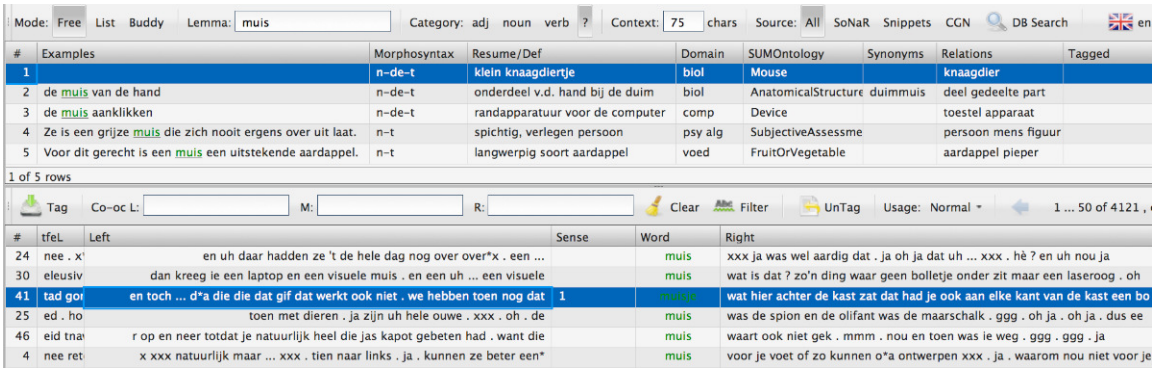

Figure 3: Annotation tool interface.

5.1 Initial balanced-sense corpus

The annotation process took about two years. In this time span, eight annotators double annotated 282,503 tokens, working 12 hours per week. As a result, 80% of the senses received 25 annotated examples or more, and 90% of the lemmas received 25 examples for each sense. The distribution of annotated examples over the different resources is 67% SoNaR, 5% CGN, and 28% web-snippets. This shows that even a 500-million-token corpus like SoNaR is not big enough to provide a balanced-sense corpus, since 28% of the examples had to be derived from the Web. Web-snippets were imported using the Snippet-tool (see Figures 4 and 5).

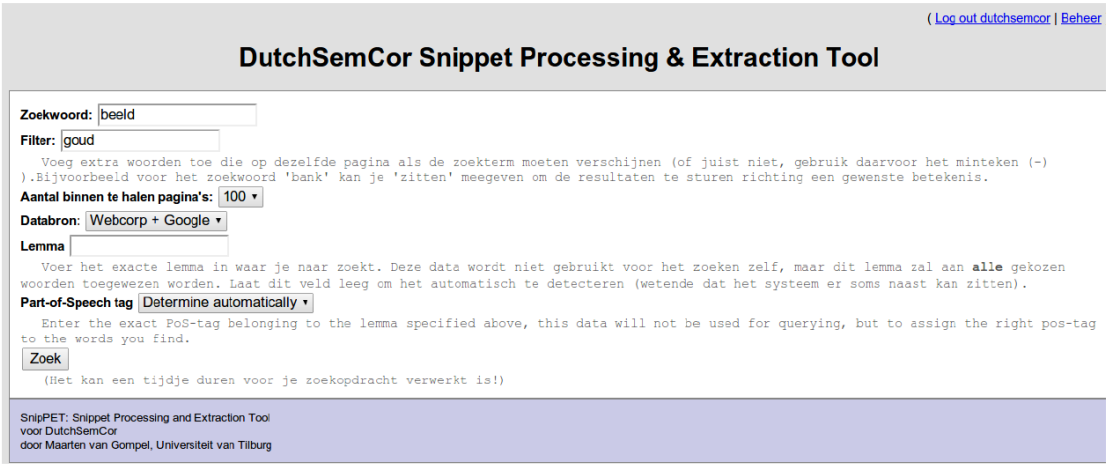

Figure 4: Screenshots of the WebSnippet tool.

Nonetheless, a small but significant portion of senses is still not well represented in the corpus even after Web search. These are mostly very rare senses belonging to specific domains or registers (e.g. one of the senses of the Dutch word *crisis* refers to a *specific critical medical state*). Nevertheless, we can conclude that we achieved a satisfactory result on the first quantitative requirement to represent all the senses of the top 2,870 most frequent and most polysemous Dutch words.

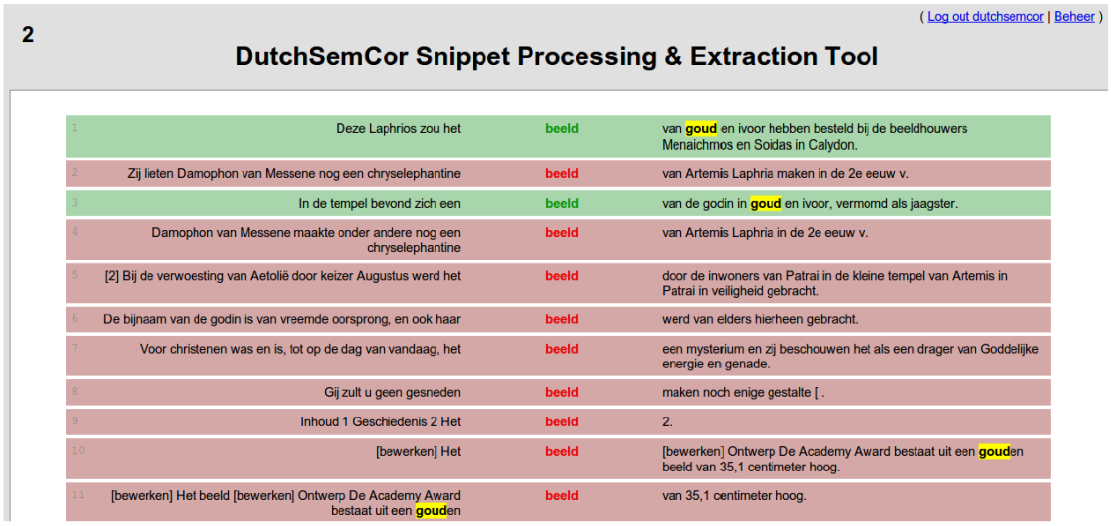

Figure 5: Screenshots of the results of the WebSnippet tool.

The average IAA for this corpus was 94%. This high IAA score can be explained by our working method: annotators did not tag all tokens presented to them, but were given the instruction to select contexts that clearly represented the senses and to avoid vague, problematic and unclear cases. This is another indication that the annotated tokens represent the senses well[6].

5.2 WSD from balanced-sense data

After creating an initial balanced-sense corpus through manual annotation, we trained and evaluated a WSD system using this data to obtain an estimation of the performance on each word. The result of this evaluation was then used to automatically conduct further annotation for weakly performing words. For this purpose, only the system **DSC-TiMBL** was used as described in Section 4. The main was that at this point of the project the DSC-TiMBL system was in a more advanced stage of development and also, as will be seen in the next section, it allowed us to easily select instances similar or different to our annotations (in terms of a similarity metric obtained from the features).

We applied a 5-fold cross validation. It was very important to test the system both for high- and low frequent senses under the same conditions. This enabled us to obtain a balanced evaluation for all senses. (Recall that in the initial annotation phase, annotators were asked to tag all senses for each word with at least 25 examples.) The folds were created at the word-sense level and not at the word level: for each word, each fold contained the same number of examples for each of its senses (randomly selected).

Since our main objective was to build a system to annotate the remainder of the corpus, we could exploit all SoNaR metadata as features. Our experiments showed, for instance, that the token identifiers of SoNaR of the annotated instances are all strong features for WSD. The effect is comparable to the one-sense-per-discourse/domain/genre heuristic (Gale et al. 1992). We can better see this effect in the next example. Consider the word **paard** in Dutch that stands for **horse** in English. As in English, this word can represent an animal or a chess piece. We found one annotation for **paard** as chess piece with this token identifier *WR-P-P-G-0000148955.p.28.s.3.w.8.* This identifier encodes the SoNaR category (WR-P-P-G), the SoNaR document (0000148955) and

6. Note that annotators could propose new senses to be added to the database or senses to be removed.

the paragraph, sentence and number of token (p.28.s.3.w.8). As we used some string matching similary measures for our DSC-TiMBL system, the system is able to detect that the token *WR-P-P-G-0000148955.p.30.s.3.w.5* for **paard** belongs to the same document as the previous occurence, and is very likely that this new token refers to the chess piece.

We ran the first evaluation for all words but focusing mainly on the nouns. The accuracy of the system for all nouns was 82.76%. From this evaluation, we selected a set of 82 lemmas performing below 80%. The output of the system for the 82 noun lemmas was validated by human annotators in three different cycles till we reached 81.62% for a total of 8,641 instances in the last evaluation round.

6. Making the corpus more balanced for context

In the second phase of the project, we tried to improve the range of contexts for the different senses. If we could annotate the full corpus, the range of contexts would be as broad as the diversity of the corpus. To minimise the effort, we thus decided to improve the WSD for the automatic annotation task by adding more examples and contexts for words that are problematic for the system. An Active Learning approach was followed to improve the classifiers for the worse performing words. Active Learning is an approach related to semi-supervised Machine Learning where the users and the system interact in several phases until the desired output or performance is reached. In our case, the system tags a set of instances with the guessed meaning, and the students correct these labels. Then the system is retrained with the new corrected data and another cycle is run. We applied the following procedure for this:

1. Select all words that perform with less than 80% accuracy on the cross-fold validation;
2. Automatically annotate the remainder of the tokens of these words using the TiMBL-WSD system;
3. From the automatically annotated tokens, we selected 50 new tokens belonging to senses that performed weakly and that had a context different from the training data. We measured this by selecting tokens with both high-confidence scores for the sense and high-distance from the k-nearest-neighbour;
4. Annotators had to annotate all the 50 tokens, i.e. they could not choose tokens that fit the senses well but had to link senses to the respective tokens;

The last point constitutes an important difference between annotation performed for the balanced-sense and the balanced-context corpus. For the former, the annotators search tokens that fit the senses, while for the latter they fit the senses to the preselected tokens. The balanced-context tokens are therefore mainly determined by the characteristics of the SoNaR-500 corpus.

The annotators were presented with 50 tokens that the system considers to belong to a 'weak' sense with high confidence. Some words have several weak senses, which results in more than 50 tokens for a word to annotate. The students independently assigned the proper senses to the tokens, without knowing the choice of the system. While annotating, they might agree or disagree with the system. In total 114,162 tokens were annotated this way. The annotators also encountered errors in lemmatisation and part-of-speech tagging, figurative and idiomatic usage and unknown senses which were marked accordingly and were excluded from the process (these represented 18% of the selected tokens).

6.1 Evaluating the extension with more contexts

We experimented with various selections of the new annotations to measure how much the WSD system will improve using the new annotations. We divided the new annotations into two groups:

Data Type	Accuracy	Num. Examples
BS	81.62	8,641
BS + LD	78.81	13,266
BS + LD_agree	85.02	11,405
BS + HD	76.24	19,055
BS + HD_agree	83.77	13,359
BS + LD_agree + HD_agree	85.33	16,123

Table 2: Evaluating the extension with more contexts.

- Low Distance[7] (LD): those with a low distance to the training instances (only marginally different contexts)
- High Distance (HD): with a high distance to the training distance (very different contexts)

We also split the new data based on the agreement of the annotators with the suggestions of the system. Considering the above divisions of the newly annotated examples, different sets were added to the initial balanced-sense (BS) corpus. We calculated the accuracy of the **DSC-TiMBL** system for the selected 82 noun lemmas trained with the different sets. Each time, the same 5-fold cross validation was carried out. The results can be seen in Table 2[8].

Interestingly, the best results are achieved using all the new training data (low- and high-distance) where the WSD system and the students agreed. Including all annotations or just low- or high-distance examples did not lead to major improvements. This proves once again that positive reinforcement (agreed upon data) works best but also that it does not matter whether the new instances are different or similar to the older training material. Apparently, our sense repository has a thorough coverage of senses which is then represented in the training data.

6.2 Optimised WSD systems on the whole balanced-context corpus

Next, we used the optimal set of annotations to finally build the final versions of the 3 different WSD systems explained above. We also defined a majority voting among the three systems that was evaluated on the same data. Table 3 shows the overall accuracy for the systems on the complete balanced-context corpus[9].

System	Acc. Nouns (%)	Acc. Verbs (%)	Acc. Adjs. (%)
DSC-TiMBL	83.97	83.44	78.64
DSC-SVM	82.69	84.93	79.03
DSC-UKB	73.04	55.84	56.36
Voting	88.65	87.60	83.06

Table 3: Evaluation of the WSD systems on the balanced-context corpus.

We can see that both DSC-TiMBL and DSC-SVM are quite similar in their performance, while DSC-UKB is slighty lower, as can be expected from an unsupervised system. The voting strategy

7. TiMBL provides the distance to the closest training instance when classifying a new instance.
8. LD stands for Low Distance instances, similar to the training data, while LD_agree are low distance instances only when the students agreed with the suggestion of the WSD. In the same way, HD and HD_agree represent High Distance instances, and the agreed subset of these.
9. We also developed a set of sense groups based on properties of synsets and relations. For instance, if two senses of the same word share the hyperonym, they are related and can be merged into a broader sense without semantic loss. Evaluation using these sense-groups can be found at the webpage of the project: http://www2.let.vu.nl/oz/cltl/dutchsemcor. Overall, the sense-groups lead to an improvement of 5% in accuracy.

outperforms the single systems. Also the results for nouns and verbs are quite high and similar for DSC-SVM and DSC-TiMBL, which is specially interesting if we consider that WSD systems usually perform lower for verbs than for nouns (probably the higher degree of polysemy of verbs can be one reason). This is not the case for DSC-UKB, where for verbs the performance drops around 18 points. The reason could be that the number of relations for verbs is lower (and with worse quality), and these relations are the key for building the graph used by DSC-UKB for the disambiguation. Finally the results for adjectives are lower in general, probably derived from the lower number of training instances for adjectives collected in the first phase.

6.3 Evaluating corpus representativeness

To test the performance of the WSD systems on the remainder of the corpus, we carried out a random evaluation. The training data was still skewed towards a balanced-sense corpus. A random selection from SoNaR would show how good these systems perform on general on the SoNaR corpus. For the random evaluation, we selected a stratified sample of lemmas for each performance range. We considered the following four ranges of accuracy based on the folded cross evaluation: [90% - 100%] , [80% - 90%] , [70% - 80%] and [60% - 70%]. From each of these performance ranges, 5 nouns, 5 verbs and 3 adjectives were randomly selected: a total of 52 lemmas. For all these lemmas, 100 untagged examples in SoNaR were automatically tagged by our system and then manually validated. Table 4 shows the results for the 3 systems and the voting heuristic.

System	Acc. Nouns (%)	Acc. Verbs (%)	Acc. Adjs. (%)
DSC-TiMBL	54.25	48.25	46.50
DSC-SVM	64.10	52.20	52.00
DSC-UKB	49.37	44.15	38.13
Voting	60.70	53.95	50.83

Table 4: Performance of our WSD systems on the random evaluation.

Clearly, results for the random evaluation are much lower than for the cross-fold validation. This shows the difference in approach between representing the senses and representing the corpus. Still, results are comparable to state-of-the-art results reported for English in Senseval/Semeval. In this case, the best system seems to be DSC-SVM, probably because it is the one that better generalizes from our previous data. We think that DSC-TiMBL could have overfitted our training data. The differences between nouns, verbs and adjectives are in line with the results reported on WSD literature.

7. Obtaining sense-probabilities

The manually annotated portion of the corpus does not exhibit sense-distributions. Mostly, the annotation was limited to 25 tokens per sense to make it balanced-sense and the extension was based on selections of 50 tokens per sense. Sense-frequencies could however be derived by automatically annotating the remainder of the corpus and assuming that the automatic annotation still reflects the true distribution. We thus applied the final WSD systems to the remainder of SoNaR and extracted the sense frequencies according to each system.

To evaluate the frequency distribution, we needed an independent sample reflecting similar distribution. Since the random sample contains only a small selection of words, a more natural sense distribution would follow from an all-words corpus. We created an all-words corpus from the part of the corpus that was kept separate from our selections (i.e. it had not been used for training purposes). This corpus consists of 23,907 tokens and represents 1,527 of our original lemmas (more than 53%).

We evaluated the three WSD systems on the all-words corpus applying 3 different baselines: the 1st sense in Cornetto, a random sense baseline and the most-frequent automatically annotated sense (MFS) by DSC-SVM[10].

System	Nouns	Verbs	Adjs.
1st sense	53.17	32.84	52.17
Random sense	29.52	24.99	32.16
Most frequent	61.20	50.76	54.62
DSC-TiMBL	55.76	37.96	49.0
DSC-SVM	64.58	45.81	55.70
DSC-UKB	56.81	31.37	35.93
Voting	66.09	45.68	52.24

Table 5: Performance of our WSD systems on the random evaluation.

The MFS performance for Dutch is similar to the results known for English. It thus seems that the MFS for Dutch according to our approach is performing equally well as a predictor. Our approach generates reasonable sense-probabilities in addition to our approach to obtain balanced-sense annotations.

The MFS baseline performs considerably higher than the 1st sense baseline for verbs (18 points) and nearly 30 points higher than the random baseline (57.54 against 28.26). We also experimented with using only high-confidence annotations but this does not lead to a significant difference. Finally, we got 6.36 points improvement by excluding the 5 most frequent verbs (auxiliary verbs)[11].

8. Project results and discussion

The DutchSemCor project resulted in numerous data sets and software tools, among which:

- 274,344 tokens for 2,874 lemmas manually annotated by two annotators with an IAA of 90% with the aim of obtaining a balanced-sense corpus
- 132,666 tokens for 1,133 lemmas, manually annotated by a single annotator but agreeing with the WSD-system for IAA 44%
- 47,797,684 automatic annotations by 3 WSD systems
- 28,080 sense groups, representing 6,903 word meanings, which improve performance by 5%
- corpora for random evaluation and all-words evaluation
- 3 WSD systems based on machine-learning
- 800,000 semantic relations between synsets derived from the annotations
- an improved version of the Cornetto database
- an annotation tool and web search tool that can be used to annotate more data
- statistics on figurative, idiomatic and collocational usage of words

10. The most-frequent sense baseline for DSC-TiMBL and DSC-UKB are performing less well.
11. Note that the corpus characteristics carried over by the token identifier in SoNaR is not useful for the all-words evaluation since the identifiers are completely different. In other words, there is no possible matching between training token identifiers and evaluation token identifiers, and the system can not make use of the one-sense-per-discourse heuristic (see the example given in Section 5.2). Likewise, the all-words evaluation can be seen as a good indication of quality of the systems for generic WSD which is different from the automatic annotation of SoNaR.

- data and statistics on phrasal verbs

Most of these results can be downloaded from the project website as open source data or can be licensed for research without a fee. The central question remains to what extent the sense-tagged corpus satisfies all 3 criteria, being: balanced-sense, balanced-context and reflecting sense-distributions. The first criterion was definitely met and was the starting point of the project. Senses that do not occur in SoNaR were retrieved using web search. Finally, a small set of senses were under-represented. We think that a balanced-sense corpus like DutchSemCor that, at the same time, represents the contexts and distributions of senses well is a unique data set.

We tried to obtain a balanced-context corpus in two steps. First, we added new contexts to weak senses and secondly we annotated the remainder of SoNaR which covers a wide range of language use. The random evaluation shows that our performance is lower than the cross-fold evaluation on the balanced-sense corpus but the results are still in line with state-of-the-art results for English. We think that future research is needed to find out whether the drop in results is due to context diversity or other facts.

Finally, the sense-probabilities were tested against an all-words corpus. Again, the results are compatible with state-of-the-art results for English. As such, we can expect that the sense-probabilities derived from DutchSemCor will also provide as strong a baseline as the MFS from SemCor is now for English. Last but not least, SoNaR provides many opportunities to differentiate these distributions over different domains and genres (McCarthy et al. 2007).

9. Conclusion

In this paper, we presented a classification of different sense-annotated corpora and described their (dis-)advantages. We proposed a method for meeting three different requirements for sense-tagged corpora. From a manually annotated seed corpus, we automatically extended the representative annotations through WSD, where we used high-confidence results and active learning for low-performing words. A small proportion of the words and word-senses will always be poorly represented, as their usage can only be found on the Web or their senses cannot be discriminated. Finally, we trained three WSD-systems using annotation data created manually and semi-automatically in the first and second phase of the project in order to extend the corpus with new tokens. Apart from cross-fold validation, we used an independent all-words corpus and a random corpus to validate the quality of the WSD system based on our lexical-sample corpus. We demonstrated the feasibility of our approach to efficiently build a balanced-sense lexical-sample corpus in a semi-automatic way that also reflects a variety of contexts and proper sense-distributions. We showed that our results are in line with state-of-the-art results for English which are mostly based on corpora that show sense-distributions or context-distributions. While our balanced-sense approach is important for modeling low frequent senses, we can still obtain good results for context-diversity and sense-probability.

In future research, we would like to further define the diversity of contexts in relation to the performance of different words in WSD systems. Especially, the rich and diverse genre and domain classification of SoNaR can be exploited to derive more precise knowledge about sense distributions in Dutch. Along the same line, the tokens annotated for figurative, metaphoric and idiomatic usage will provide valuable data to research. Finally, we will further experiment with different behaviors of supervised and unsupervised systems by inserting sense-probabilities assigned by the supervised systems into the graphs of the unsupervised system. We hope to integrate the learned data in a system that is more robust to changes of genre and domain.

References

Agirre, Eneko and Aitor Soroa (2009), Personalizing pagerank for word sense disambiguation, *Proceedings of the 12th Conference of the European Chapter of the Association for Computational*

Linguistics (EACL '09), pp. 33–41.

Agirre, Eneko and Philip Edmonds (2006), *Word Sense Disambiguation: Algorithms and Applications*, Text, speech and language technology, Springer, Dordrecht, NE.

Aha, David W., Dennis Kibler, and Marc K. Albert (1991), Instance-based learning algorithms, *Machine Learning* **6**, pp. 37–66, Kluwer Academic Publishers.

Bruce, Rebecca F. and Janyce Wiebe (1994), Word-sense disambiguation using decomposable models, *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL '94)*, pp. 139–146.

Cortes, Corinna and Vladimir Vapnik (1995), Support-vector networks, *Machine Learning*, pp. 273–297.

Daelemans, Walter, Jakub Zavrel, Ko Van der Sloot, and Antal Van den Bosch (2007), TiMBL: Tilburg Memory Based Learner, version 6.1, reference guide, *Technical report*, ILK Research Group Technical Report Series no. 07-07.

Fellbaum, Christiane (1998), *WordNet: An Electronic Lexical Database*, Bradford Books.

Gale, William A., Kenneth W. Church, and David Yarowsky (1992), One sense per discourse, *In DARPA Speech and Natural Language Workshop.*

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll (2007), Unsupervised acquisition of predominant word senses, *Computational Linguistics* **33** (4), pp. 553–590, Massachusetts Institute of Technology.

Miller, George A., Claudia Leacock, Randee Tengi, and Ross T. Bunker (1993), A semantic concordance, *Proceedings of the workshop on Human Language Technology (HLT '93)*, pp. 303–308.

Mooney, Raymond J. (1996), Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-96)*, pp. 82–91.

Ng, Hwee Tou and Hian Beng Lee (1996), Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach, *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL '96)*, pp. 40–47.

Ng, Hwee Tou and Hian Beng Lee (1997), *DSO Corpus of Sense-Tagged English*, Linguistic Data Consortium, Philadelphia.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste, and Ineke Schuurman (2013), The construction of a 500-million-word reference corpus of contemporary written Dutch, *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, Springer Verlag, chapter 13.

Vossen, Piek, Katja Hofmann, Maarten de Rijke, Erik Tjong, Kim Sang, and Koen Deschacht (2007), The Cornetto database: Architecture and user-scenarios, *in* Moens, M. F., T. Tuytelaars, and A. P. de Vries, editors, *Proceedings DIR 2007*, pp. 89–96.