Computer assisted semantic annotation in the DutchSemCor project

Attila Görög and Piek Vossen
Computational Lexicology & Terminology Lab
Vrije Universiteit Amsterdam

De Boelelaan 1105
1081 HV Amsterdam
The Netherlands

E-mail: a.gorog@let.vu.nl, p.vossen@let.vu.nl

Abstract

The goal of this paper is to describe the annotation protocols and the Semantic Annotation Tool (SAT) used in the DutchSemCor project. The DutchSemCor project is aiming at aligning the Cornetto lexical database with the Dutch language corpus SoNaR. 250K corpus occurrences of the 3,000 most frequent and most ambiguous Dutch nouns, adjectives and verbs are being annotated manually using the SAT. This data is then used for bootstrapping 750K extra occurrences which in turn will be checked manually. Our main focus in this paper is the methodology applied in the project to attain the envisaged Inter-annotator Agreement (IA) of ≥80%. We will also discuss one of the main objectives of DutchSemCor i.e. to provide semantically annotated language data with high scores for quantity, quality and diversity. Sample data with high scores for these three features can yield better results for co-training WSD systems. Finally, we will take a brief look at our annotation tool.

1. Introduction

The importance of semantically annotated corpora for Word Sense Disambiguation (WSD) has been underlined in various research projects in the past decade. The numerous SENSEVAL-tasks produced interesting data for the evaluation of WSD systems and provided a theoretical background for the creation of semantically annotated corpus material. Supervised and unsupervised methods to decipher meaning have been extensively tested and described, and the results have been compared to gold standards. One subject, however, gained only minor attention within the framework of WSD namely the actual process of semantic annotation by human "taggers" as well as the tools and methodology applied in the different projects.

In what follows, we will give an account of the DutchSemCor project, the methodology we have been using for the analysis of annotations and finally, the Semantic Annotation Tool (SAT) which had been developed for the computer assisted semantic tagging of corpus material. First, we will set forth the aims and purposes of the DutchSemCor project, a collaboration project between three Dutch universities (Section 1). In the project, manual tagging is combined with supervised methods and a unique methodology is applied to reach optimal scores for quantity, diversity and quality of the manually annotated data. These three scores are important for optimal co-training of our WSD-systems (Section 2). Finally, we will introduce the SAT used for the manual annotation task (Section 3).

2. The DutchSemCor project

Most NLP applications require large sense-tagged corpora along with lexical databases to reach satisfactory results in WSD tasks such as machine translation, question & answering, summarization and terminology extraction. The number of English language resources have increased in the past years, the data scarceness for other languages, however, is more than obvious.

The situation is similar for the Dutch language: scarceness of semantically annotated corpus material to train WSD machines. In order to overcome the data bottleneck the DutchSemCor project is aiming to deliver a one-million word Dutch corpus that is fully sense-tagged with senses and domain tags from the Cornetto lexical database (Vossen 2006 and Vossen et al. 2007, 2008). 250K examples of this corpus are being manually tagged. The remainder will be automatically tagged using three different WSD systems and will be validated by human annotators. The corpus data is based on existing corpus material collected in the projects CGN (Eerten, 2007), D-CoI and SoNaR (Oostdijk et al., 2008). These corpora have already been parsed and tagged in previous projects and will be extended where necessary in order to find sufficient examples for different word senses that are less frequent and do not appear in the above corpora. When writing this essay, our project is in a preliminary phase, we have currently begun with the annotation of our corpus material for Dutch nouns.

3. General methodology

In this section we will describe the different phases of the annotation project (a combination of manual and automatic techniques) followed by a short overview of the different phases of the manual annotation process.

3.1 A combination of manual and automatic annotation

We are using a mixture of automatic and manual tagging procedures. The envisaged corpus of 1 million tokens is split into two parts that are handled in different ways. The first part of about 250,000 tokens is being annotated at the moment in a traditional way: on average 25 examples of each meaning of 3,000 most frequent and most polyseme words of the Dutch language (65% nouns, 23% verbs and 12% adjectives) are analyzed and tagged by a group of 8 human annotators. This tagging is supported by a knowledge-rich tagging system (see next section) that does not rely on training examples. We are counting on an average of 3.4 senses per word (based on data in the Cornetto database).

The second part of the corpus will cover 750,000 tokens, adding another 75 examples for each word meaning. The coverage of the corpus is partly based on the remainders of the general corpora used, and partly on the necessity to find sufficient examples for each meaning of the selected words. This second part of the corpus will be tagged automatically at a later stage using tagging systems that are trained by the manually tagged data acquired so far and any other data that can be used (bootstrapping). The manual tagging in the second phase then involves validating the automatic assignments by a human annotator. This means that we focus on those cases where the confidence of the system is low and different systems disagree, as in active learning or co-training methods. Note that we can also group word occurrences based on their estimated meanings and compare the different contexts in which they occur. If there are insufficient examples of a word in a particular meaning in the corpus, sampling techniques can be used to find additional examples of the word in its context, e.g. on the Web or in large textual corpora.

3.2 Different phases of manual annotation

In what follows we will discuss the process of manual tagging. Already after the first annotation sequences of our project, it has become obvious that high agreement scores and reasonable quality of annotated material can only be reached if the annotators have a clear and unanimous perception of the different senses of a lemma. For this reason, we have introduced project meetings at a very early stage of our project. In these meetings, involving the 8 annotators and the two coordinators we reflect on problems of different origins (possible mistakes in the lexical database, difficult sense distinctions, senses that are not represented in the corpus, etc). Also, we discuss co-occurrence strategies to find word meanings directly in the corpus or on the Internet as well as to group examples and to discover figurative and idiomatic uses. Another purpose of the discussions is to gain insight into the peculiarities of the Dutch language and to teach annotators test their language instincts using different word-meaning tests (e.g. zeugma, cross readings etc).

In order to reach an Inter-annotator Agreement of minimum 80%, we implement the following working cycle divided into three different phases: pre-processing the Cornetto data, preliminary discussion (Preparatory phase); manual annotation sessions (Annotation phase); Post-editing the Cornetto data (Editorial phase).

3.2.1 Pre-processing the Cornetto data

Before the preliminary discussion, the Cornetto data needs to be inspected by the coordinator of the project and if necessary the entries need to be corrected. Also a word list is to be prepared. Every two weeks a new word list of approx. 200 words are processed by 8 annotators (4 couples). The editing process consists of the following main tasks: delineation of word meanings, verifying the alignment between LUs and Synsets, splitting, merging or removing senses, if necessary creating new senses, adding morpho-syntactic/ semantic information, adding examples, synonyms, etc.

3.2.2 Preliminary discussion

We hold one meeting of two hours per week. An important part of these meetings is the preliminary discussion of ‘new words’. During this preliminary discussion, the coordinator of the annotation project points out possible difficulties based on data from the Cornetto lexical database. The aim is to prepare annotators for certain pitfalls common in human WSD tasks and to suggest methods to overcome these difficulties.

3.2.3 Manual annotation 1

Two annotators (A1 + A2) receive the same words and the same KWIC index examples of the reference corpus to annotate. Note that the annotators are free to choose or ignore certain examples. (The annotation tool restricts the number of examples per sense otherwise there would be too little overlap between the tagged instances). The resulting overlap between the annotated occurrences can be divided into two groups. One group contains the tokens for which an agreement has been reached (see figure 1 – **Group 1**). These examples are identically tagged between the two annotators and need not further be discussed. The other group (see figure 1 – **Group 2**) are those occurrences which have been tagged differently by the two annotators. During the 1st discussion we will look at these examples. It is important to account for the differences and in some cases the sense division of the given Cornetto entry needs to be changed.

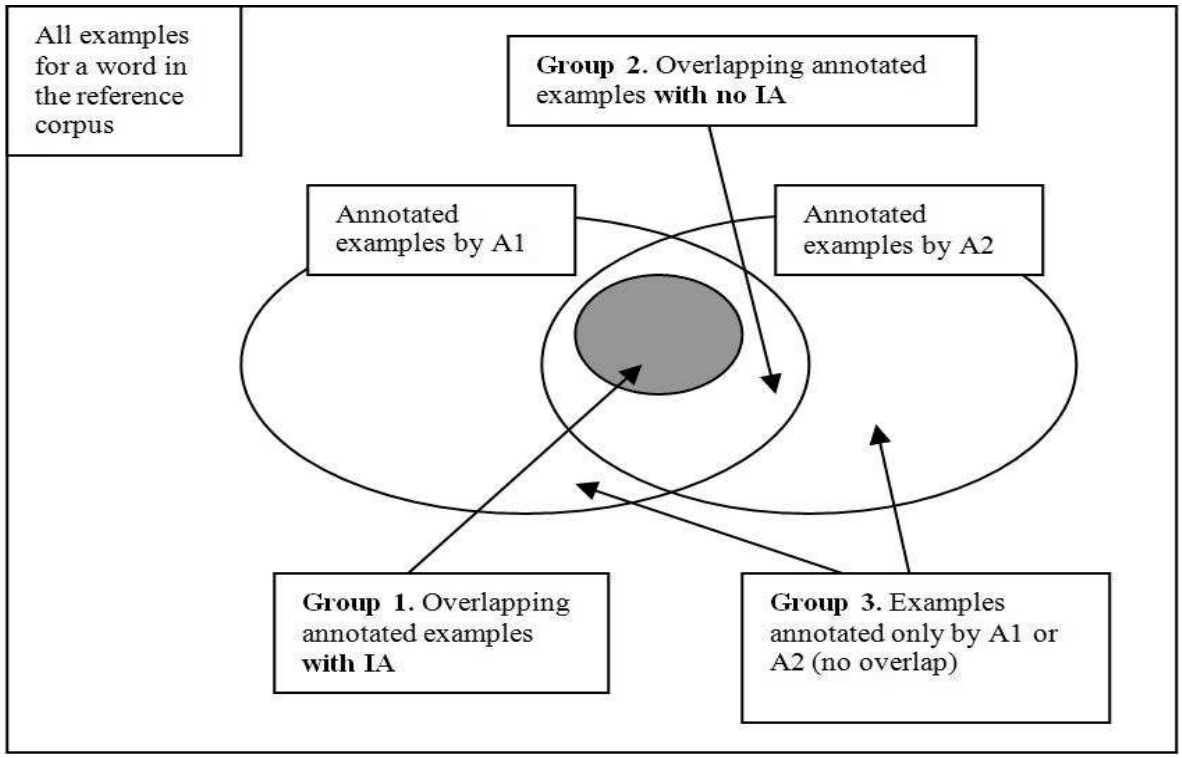


Figure 1: Results of the first sequence of manual annotation

3.2.4 Manual annotation 2

After the 1st discussion, the examples in **Group 2** are annotated again by the two annotators. Most of the examples of **Group 2** are annotated identically in the second round (due to the previous discussions and clarifications of word meanings) increasing this way the overall IA. The examples of **Group 3** will be exchanged between the two annotators (these are the examples which have been annotated by only one of the two annotators) and a 2nd discussion follows (**see 3.2.3**). The result of this procedure is that an IA of minimum **80%** is reached for the three groups mentioned above at the end of the second week**.**

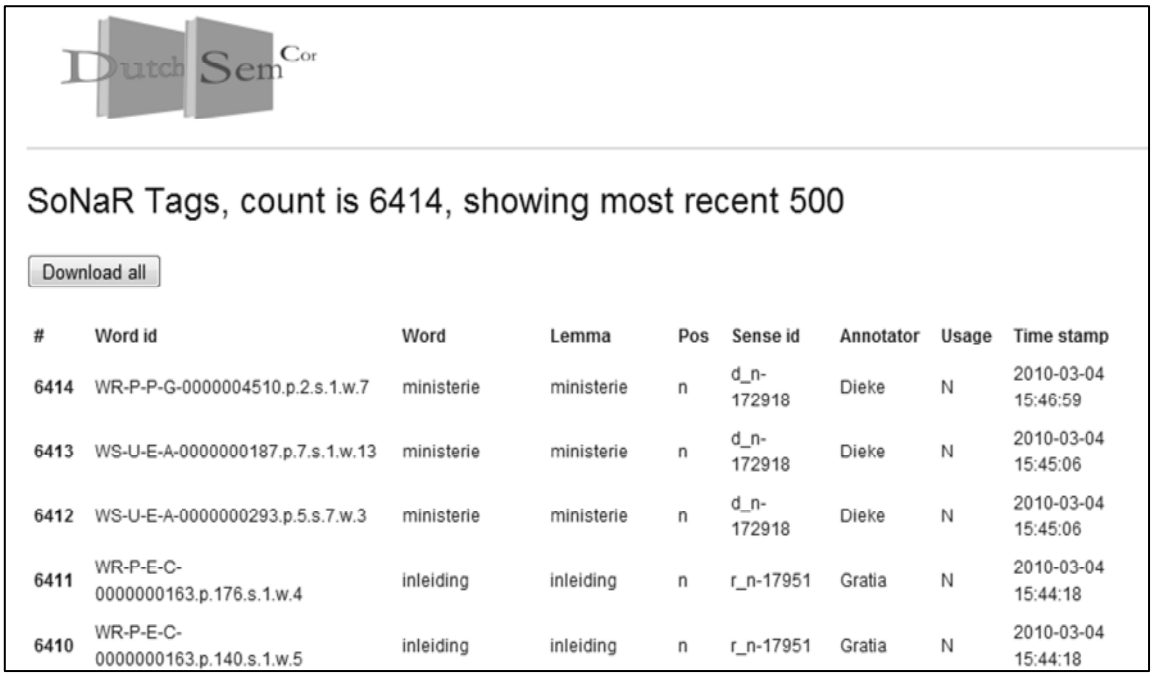
3.2.5 Post-editing the Cornetto data

Based on the annotated occurrences in the corpus, our aim is to, if necessary, correct senses or create new additional senses in the Cornetto lexical database. This happens using the following steps:

1. **Clustering senses based on corpus data** à using lexical-contextual clues and syntactic patterns within paragraph.
2. **Choose 'prototypical sense'** à from cluster (based on frequency and intuition).
3. **Determine the different 'shifts'** à This shows the meaning changes from 'prototype sense' to other senses (metaphor, metonymy etc.) and the sense divisions inside a sense inventory.
4. **If necessary *merge/ split* senses**

Summary of the annotation process:

1.	**Pre-processing Cornetto data**
	Preliminary discussion
2.	**Annotation phase 1** 3 groups of examples: **Goup 1** = overlap, IA **Group 2** = overlap, no IA **Group 3** = no overlap, no IA **Discussion 1** **Group 1** = OK; discuss **Group 2**
3.	**Annotation phase 2** re-annotation **Group 2** (reaching IA) annotation **Group 3** **Discussion 2** **Group 1+2** = OK; discuss **Group 3** re-annotation **Group 3** **Group 1+2+3** = IA ≥ 80%
4.	**post-editing Cornetto entries**



SoNaR Tags, count is 6414, showing most recent 500

Download all

#	Word id	Word	Lemma	Pos	Sense id	Annotator	Usage	Time stamp
6414	WR-P-P-G-0000004510.p.2.s.1.w.7	ministerie	ministerie	n	d_n-172918	Dieke	N	2010-03-04 15:46:59
6413	WS-U-E-A-0000000187.p.7.s.1.w.13	ministerie	ministerie	n	d_n-172918	Dieke	N	2010-03-04 15:45:06
6412	WS-U-E-A-0000000293.p.5.s.7.w.3	ministerie	ministerie	n	d_n-172918	Dieke	N	2010-03-04 15:45:06
6411	WR-P-E-C-0000000163.p.176.s.1.w.4	inleiding	inleiding	n	r_n-17951	Gratia	N	2010-03-04 15:44:18
6410	WR-P-E-C-0000000163.p.140.s.1.w.5	inleiding	inleiding	n	r_n-17951	Gratia	N	2010-03-04 15:44:18

Figure 2: Screenshot of the log-file

3.3 Quantity, diversity and quality of data

In previous projects such as OntoNotes (Sameer and Nianwen, 2009) similar cycli have been used as the one mentioned above in order to reach high IA scores. To our knowledge, no further criteria have been applied in these projects. Our aim is to not only obtain an IA score of minimum 80% but also to deliver a large corpus which is sufficiently diverse in terms of syntactic and semantic patterns.

Based on a detailed log-file, annotation results are evaluated and discussed with the annotators. Each tagged sentence and every annotator action is recorded in a log-file. Since every corpus fragment receives an ID it is possible to analyze the quality, diversity and quantity of the tagged instances. (See Figure 2 for an example of the log-file).

We are trying to reach high diversity by implementing different filters which make use of constituency patterns, semantic roles, collocational information, domain labels etc. (for automatic pre-labelling of paragraphs with domain labels see 3.4).

Finally, the IA-score is our quality measurement and is very useful for the different discussions with the annotators. Low agreement usually means difficulties either in linking the right examples to the existing Cornetto senses or problems with the sense divisions of Cornetto itself. This latter will need to be corrected by the coordinator of the annotation project.

This way, we not only guarantee rich and interesting data for purposes of linguistic research but also a semantic corpus with optimal variation for machine learning. Text fragments with a great syntactic and semantic diversity can better serve WSD techniques and yield better results when used for bootstrapping (see also Ng, 1997).

The log-file is converted into a feature table by a *log-analyzer* (a tool developed by the Vrije Universiteit Amsterdam). The table contains different information and scores for the above mentioned features (see Figure 3).

ervaring								
	Nr of annotators	2						
	Nr of tokens	143						
	Annotator	Nr tokens	Nr annos	Nr unique	Nr overlap	Nr Agree	Nr Disagree	Nr Partial
	Gratia	81	81	68	13	10	3	0
	Wilma	75	75	62	13	10	3	0
	Gratia							
Senses	Nr tokens	Nr unique	Nr overlap	Nr Agree	Nr Disagree	Nr Partial		
r_n-12605	25	21	4	4	0	0		
r_n-12604	28	25	3	3	0	0		
r_n-12603	28	25	3	3	0	0		
	Wilma							
Senses	Nr tokens	Nr unique	Nr overlap	Nr Agree	Nr Disagree	Nr Partial		
r_n-12605	25	21	4	4	0	0		
r_n-12604	24	21	3	3	0	0		
r_n-12603	26	23	3	3	0	0		
		IAA exact	IAA weak	Diversity	Start time	End time		
	Gratia	76	76		1-3-2010 20:28	1-3-2010 20:40		
	Wilma	76	76		2-3-2010 11:30	2-3-2010 11:40		

Figure 3: output of the *log-analyzer*

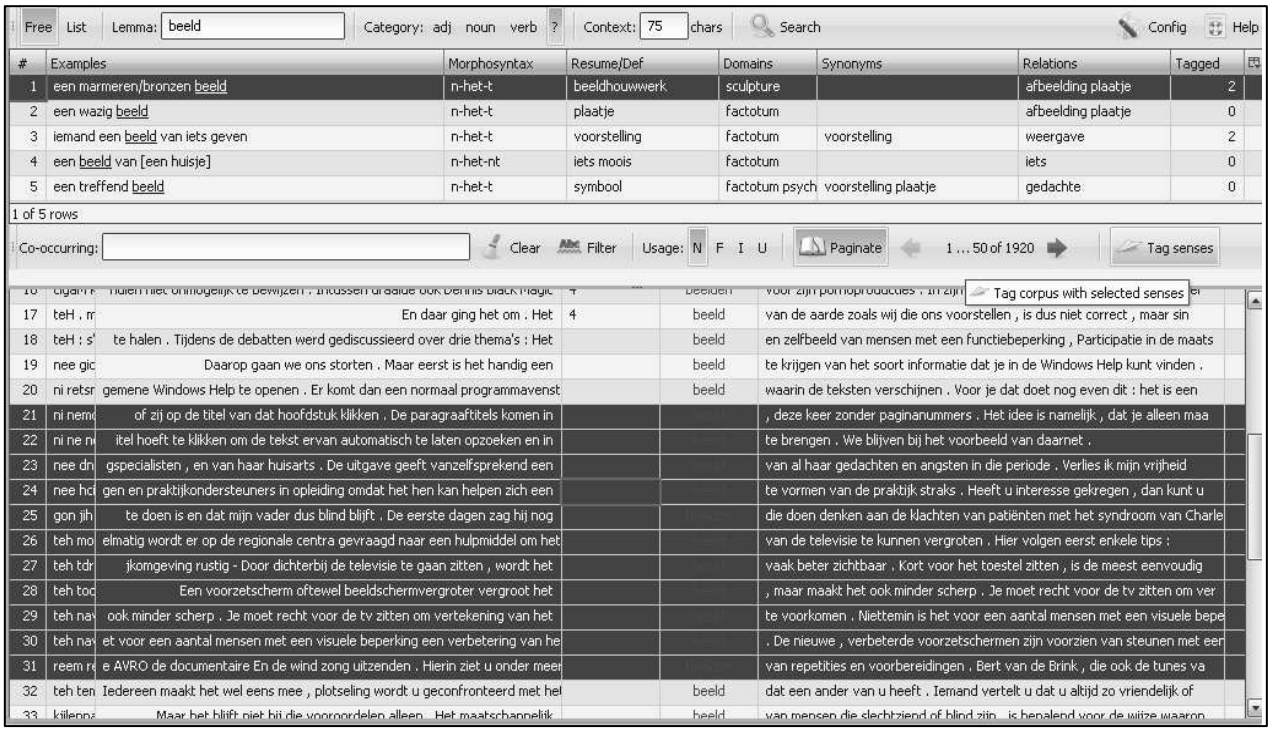


Figure 4: Linking senses of the Dutch word 'beeld' (*Eng.* 'figure, image') with corpus examples in the SAT

4. Semantic Annotation Tool (SAT)

4.1 Different features of the SAT

Our semantic annotation tool provides human annotators with an ergonomic and easy to use web-based environment in which an optimal result can be reached for computer assisted semantic annotation. The SAT gives access to the Cornetto database and to text fragments from the reference corpus. Cornetto is a semantically rich lexical database which contains the Dutch WordNet, the RBN (Referentie Bestand Nederlands, a Dutch lexicon developped by the Vrije Universiteit Amsterdam) and is also enriched with other semantic layers (WordNet Domains and the SUMO ontology). Based on different types of information (definitions, examples, grammatical and semantic information), human annotators are asked to link corpus examples to Cornetto-senses (Figure 4).

Figure 5: Pop-up window for extra context in the SAT

For the purpose of targeted tagging, all occurrences of a word are displayed in a sortable KWIC-index (targeted tagging) and interfaced with the meaning specification in the Cornetto database. Special measures are taken to detect and exclude idiomatic usages of words from the retrieved text. In case these multi-word units cannot be excluded automatically, annotators mark them (I = Idiom). Furthermore, if a certain meaning of a word found in the corpus does not occur in Cornetto, it is labeled by the human annotator as a new word meaning (U= Unknown) and added to the database during an editorial round. Similarly, the sense-annotation tool supports labeling figurative usage and metonymic usage (F= Figurative).

The tool is built in a way that only necessary information is presented at once in the different windows but standing with the cursor on the relevant data, more information is provided for each field (e.g. more context, more synonyms, hyponymy/ hypernymy relations, domain labels etc.) (Figure 5). This way, the annotator is able to decide which extra information he/she needs in order to correctly assign senses to different occurrences.

It is also possible to group corpus examples according to different criteria (words left or right to the target word) and to search examples using different word-clues (e.g.: multi word search). If the number of text fragments is insufficient, users can also launch a web-search enriching this way the internal corpus with new text fragments.

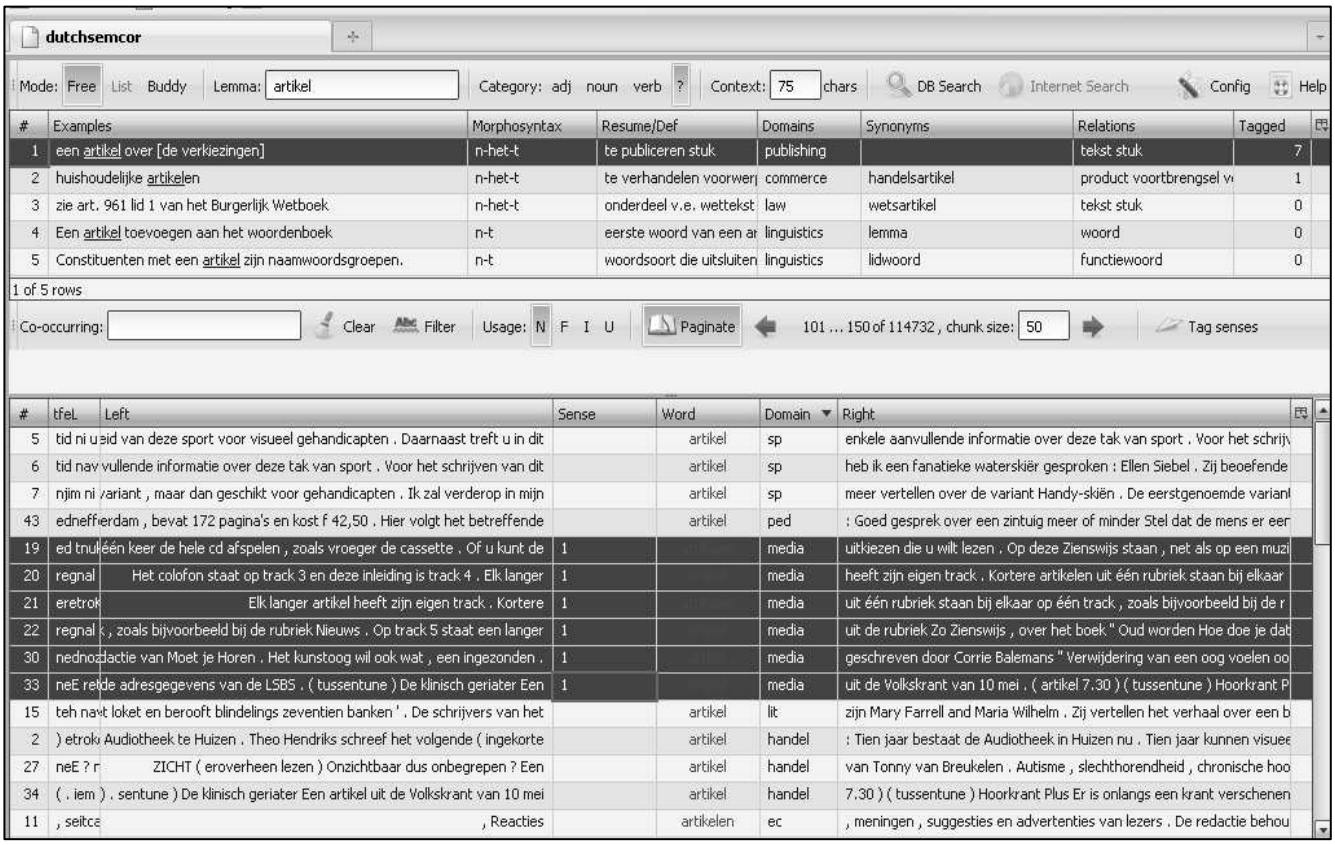


Figure 6: Automatically generated domain labels for occurrences of the Dutch word 'artikel' (*Eng.*'article') in the corpus

4.2 Using the classifier to pre-label paragraphs with domain labels

As we have mentioned in the previous section, the SAT contains different filters by which the user can re-group, analyze or restrict data in several ways. One of the filters provided in the tool is a classifier which automatically assigns domain labels to corpus occurrences. The resulting data can be sorted according to the domains facilitating this way the matching of corpus examples to Cornetto senses (which themselves are marked by domain labels).

The classification engine, a product of Irion technologies (http://www.irion.nl/) allows the user to train a classifier by giving it a set of paragraphs with classes. The classifier can then assign these classes to unseen paragraphs. For the classes a list of WordNet Domain labels is used and mapped onto the Cornetto senses (Figure 5). When classifying a corpus fragment, it will compare the signature of the incoming text fragment with the paragraphs in the training set an extract a score for the categories of the most similar paragraph. The domain labels can be organized hierarchically and the system can assign more than one label to a fragment. The system provides many options and tools to evaluate the quality of the classifier and to give feedback and suggestions to improve it.

5. Conclusion

Semantic annotation of text corpora is a task requiring enormous intellectual effort. The DutchSemCor project is aiming at the human annotation of 250K words and the human validation of a 750-word automatically sense tagged corpus. To achieve such numbers, the implementation of a user-friendly and semantically rich annotation tool is indispensable. Before developing semantic annotation software, it is important to plan the different phases and steps of the annotation project, the evaluation of annotations, the scoring etc. in one word the methodology. The right methodological approach and a user-friendly tool with an intelligent design are necessary assets for successful semantic annotation.

(For a first impression of the SAT, please visit: http://cornetto.science.uva.nl:8080/dutchsemcor/)

6. References

Agirre, E., Stevenson, M. (2006). Knowledge sources for WSD. In *Word Sense Disambiguation: Algorithms and Applications.* New York, NY : Springer, pp. 217--251.

Eerten, L. (2007). Over het Corpus Gesproken Nederlands. In *Nederlandse Taalkunde*, 12 (3) pp. 194--215.

Kilgarriff, A. (2006). Word senses. In *Word Sense Disambiguation: Algorithms and Applications*. New York, NY : Springer, pp. 29--46.

Mihalcea, R. (2002) Bootstrapping large sense tagged corpora. In *Proceedings of the 3rd International Conference on Language Resources and Evaluations* (LREC 2002), Las Palmas, Spain.

Mihalcea, R. (2004). Co-training and self-training for word sense disambiguation. In *Proceedings of the 8th Conference on Computational Natural Language Learning* CoNLL, Boston, MA, 33--40.

Navigli, R. (2009). Word Sense Disambiguation: a Survey. In *ACM Computing Surveys*, 41(2), ACM Press. pp. 1--69.

Ng, H. T., (1997). Getting serious about word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, U.S.A., 1--7.

Oostdijk, N. et al. (2008). From D-Coi to SoNaR: A reference corpus for Dutch. In: *Proceedings on the sixth international Conference on Language Resources and Evaluation* (LREC 2008), Marrakech, Marocco.

Palmer, M., Ng, H. T., Dang, H. T. (2006). Evaluation of WSD systems. In *Word Sense Disambiguation: Algorithms and Applications.* New York, NY : Springer, pp. 75--106.

Pianta, E., Bentivogli, L. (2003). Translation as Annotation, In *Proceedings of the AI*IA 2003 Workshop `Topics and Perspectives of Natural Language Processing in Italy`*, Pisa, Italy.

Sameer, S. P., Nianwen, X. (2009). OntoNotes: the 90% solution, In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, Association for Computational Linguistics, Boulder, Colorado.

Vossen, P. (2006). Cornetto: Een lexicaal-semantische database voor taaltechnologie, *Dixit Special Issue*, Stevin.

Vossen, P. et al. (2007). The Cornetto Database: Architecture and User-Scenarios. In *DIR*.pp.89--96.

Vossen, P. et al. (2008). Integrating Lexical Units, Synsets, and Ontology in the Cornetto Database. In: *Proceedings on the sixth international Conference on Language Resources and Evaluation* (LREC 2008), Marrakech, Marocco.