# A platform for collaborative semantic annotation

**Valerio Basile** and **Johan Bos** and **Kilian Evang** and **Noortje Venhuizen**
{v.basile,johan.bos,k.evang,n.j.venhuizen}@rug.nl
Center for Language and Cognition Groningen (CLCG)
University of Groningen, The Netherlands

## Abstract

Data-driven approaches in computational semantics are not common because there are only few semantically annotated resources available. We are building a large corpus of public-domain English texts and annotate them semi-automatically with syntactic structures (derivations in Combinatory Categorial Grammar) and semantic representations (Discourse Representation Structures), including events, thematic roles, named entities, anaphora, scope, and rhetorical structure. We have created a wiki-like Web-based platform on which a crowd of expert annotators (i.e. linguists) can log in and adjust linguistic analyses in real time, at various levels of analysis, such as boundaries (tokens, sentences) and tags (part of speech, lexical categories). The demo will illustrate the different features of the platform, including navigation, visualization and editing.

## 1 Introduction

Data-driven approaches in computational semantics are still rare because there are not many large annotated resources that provide empirical information about anaphora, presupposition, scope, events, tense, thematic roles, named entities, word senses, ellipsis, discourse segmentation and rhetorical relations in a single formalism. This is not surprising, as it is challenging and time-consuming to create such a resource from scratch.

Nevertheless, our objective is to develop a large annotated corpus of Discourse Representation Structures (Kamp and Reyle, 1993), comprising most of the aforementioned phenomena: the Groningen Meaning Bank (GMB). We aim to reach this goal by:

1. Providing a wiki-like platform supporting collaborative annotation efforts;

2. Employing state-of-the-art NLP software for bootstrapping semantic analysis;

3. Giving real-time feedback of annotation adjustments in their resulting syntactic and semantic analysis;

4. Ensuring kerfuffle-free dissemination of our semantic resource by considering only public-domain texts for annotation.

We have developed the wiki-like platform from scratch simply because existing annotation systems, such as GATE (Dowman et al., 2005), NITE (Carletta et al., 2003), or UIMA (Hahn et al., 2007), do not offer the functionality required for deep semantic annotation combined with crowdsourcing.

In this description of our platform, we motivate our choice of data and explain how we manage it (Section 2), we describe the complete toolchain of NLP components employed in the annotation-feedback process (Section 3), and the Web-based interface itself is introduced, describing how linguists can adjust boundaries of tokens and sentences, and revise tags of named entities, parts of speech and lexical categories (Section 4).

## 2 Data

The goal of the Groningen Meaning Bank is to provide a widely available corpus of texts, with deep semantic annotations. The GMB only comprises texts from the public domain, whose distribution isn't subject to copyright restrictions. Moreover, we include texts from various genres and sources, resulting in a rich, comprehensive



*Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 92–96,
Avignon, France, April 23 - 27 2012. 

corpus appropriate for use in various disciplines within NLP.

The documents in the current version of the GMB are all in English and originate from four main sources: (i) *Voice of America* (VOA), an online newspaper published by the US Federal Government; (ii) the *Manually Annotated Sub-Corpus* (MASC) from the Open American National Corpus (Ide et al., 2010); (iii) country descriptions from the *CIA World Factbook* (CIA) (Central Intelligence Agency, 2006), in particular the Background and Economy sections, and (iv) a collection of Aesop's fables (AF). All these documents are in the public domain and are thus redistributable, unlike for example the WSJ data used in the Penn Treebank (Miltsakaki et al., 2004).

Each document is stored with a separate file containing metadata. This may include the language the text is written in, the genre, date of publication, source, title, and terms of use of the document. This metadata is stored as a simple feature-value list.

The documents in the GMB are categorized with different statuses. Initially, newly added documents are labeled as *uncategorized*. As we manually review them, they are relabeled as either *accepted* (document will be part of the next stable version, which will be released in regular intervals), *postponed* (there is some difficulty with the document that can possibly be solved in the future) or *rejected* (something is wrong with the document form, i.e., character encoding, or with the content, e.g., it contains offensive material).

Currently, the GMB comprises 70K English text documents (Table 1), corresponding to 1,3 million sentences and 31,5 million tokens.

Table 1: Documents in the GMB, as of March 5, 2012

| Documents | VOA | MASC | CIA | AF | All |
|---|---|---|---|---|---|
| Accepted | 4,651 | 34 | 515 | 0 | 5,200 |
| Uncategorized | 61,090 | 0 | 0 | 834 | 61,924 |
| Postponed | 2,397 | 339 | 3 | 1 | 2,740 |
| Rejected | 184 | 27 | 4 | 0 | 215 |
| Total | 68,322 | 400 | 522 | 835 | 70,079 |

## 3 The NLP Toolchain

The process of building the Groningen Meaning Bank takes place in a bootstrapping fashion. A chain of software is run, taking the raw text documents as input. The output of this automatic process is in the form of several layers of stand-off annotations, i.e., files with links to the original, raw documents.

We employ a chain of NLP components that carry out, respectively, tokenization and sentence boundary detection, POS tagging, lemmatization, named entity recognition, supertagging, parsing using the formalism of Combinatory Categorial Grammar (Steedman, 2001), and semantic and discourse analysis using the framework of Discourse Representation Theory (DRT) (Kamp and Reyle, 1993) with rhetorical relations (Asher, 1993).

The lemmatizer used is morpha (Minnen et al., 2001), the other steps are carried out by the C&C tools (Curran et al., 2007) and Boxer (Bos, 2008).

### 3.1 Bits of Wisdom

After each step in the toolchain, the intermediate result may be automatically adjusted by auxiliary components that apply annotations provided by expert users or other sources. These annotations are represented as "Bits of Wisdom" (BOWs): tuples of information regarding, for example, token and sentence boundaries, tags, word senses or discourse relations. They are stored in a MySQL database and can originate from three different sources: (i) explicit annotation changes made by experts using the Explorer Web interface (see Section 4); (ii) an annotation game played by non-experts, similar to 'games with a purpose' like Phrase Detectives (Chamberlain et al., 2008) and Jeux de Mots (Artignan et al., 2009); and (iii) external NLP tools (e.g. for word sense disambiguation or co-reference resolution).

Since BOWs come from various sources, they may contradict each other. In such cases, a judge component resolves the conflict, currently by preferring the most recent expert BOW. Future work will involve the application of different judging techniques.

### 3.2 Processing Cycle

The widely known open-source tool *GNU make* is used to orchestrate the toolchain while avoiding unnecessary reprocessing. The need to rerun the toolchain for a document arises in three situations: a new BOW for that document is available; a new, improved version of one of the components is available; or reprocessing is forced by a user via the "reprocess" button in the Web interface. A continually running program, the 'updat-


Figure 1: A screenshot of the web interface, displaying a tokenised document.

ing daemon', is responsible for calling *make* for the right document at the right time. It checks the database for new BOWs or manual reprocessing requests in very short intervals to ensure immediate response to changes experts make via the Web interface. It also updates and rebuilds the components in longer intervals and continuously loops through all documents, remaking them with the newest versions of the components. The number of *make* processes that can run in parallel is configurable; standard techniques of concurrent programming are used to prevent more than one *make* process from working simultaneously on the same document.

## 4 The Expert Interface

We developed a wiki-like Web interface, called the GMB Explorer, that provides users access to the Groningen Meaning Bank. It fulfills three main functions: navigation and search through the documents, visualization of the different levels of annotation, and manual correction of the annotations. We will discuss these functions below.

### 4.1 Navigation and Search

The GMB Explorer allows navigation through the documents of the GMB with their stand-off annotations (Figure 1). The default order of documents is based on their size in terms of number of tokens. It is possible to apply filters to restrict the set of documents to be shown: showing only documents from a specific subcorpus, or specifically showing documents with/without warnings generated by the NLP toolchain.

The Explorer interface comes with a built-in search engine. It allows users to pose single- or multi-word queries. The search results can then be restricted further by looking for a specific lexical category or part of speech. A more advanced search system that is based on a *semantic lexicon* with lexical information about all levels of annotation is currently under development.

### 4.2 Visualization

The different visualization options for a document are placed in tabs: each tab corresponds to a specific layer of annotation or additional information. Besides the raw document text, users can view its tokenized version, an interactive derivation tree per sentence, and the semantic representation of the entire discourse in graphical DRS format. There are three further tabs in the Explorer: a tab containing the warnings produced by the NLP pipeline (if any), one containing the Bits of Wisdom that have been collected for the document, and a tab with the document metadata.

The *sentences* view allows the user to show or hide sub-trees per sentence and additional information such as POS-tags, word senses, supertags and partial, unresolved semantics. The derivations are shown using the CCG notation, generated by XSLT stylesheets applied to Boxer's XML output. An example is shown in Figure 2.

The *discourse* view shows a fully resolved semantic representation in the form of a DRS with

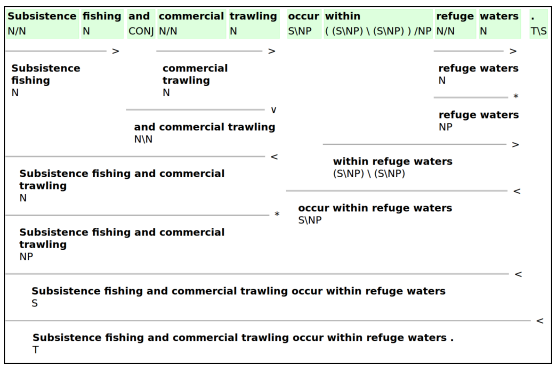


Figure 2: An example of a CCG derivation as shown in GMB Explorer.

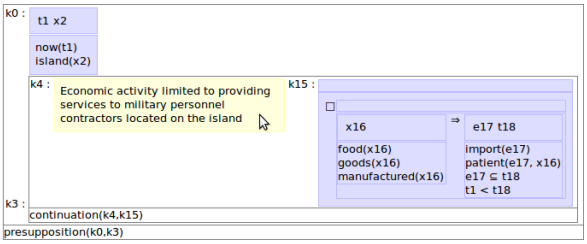


Figure 3: An example of the semantic representations in the GMB, with DRSs representing discourse units.

rhetorical relations. Clicking on discourse units switches the visualization between text and semantic representation. Figure 3 shows how DRSs are visualized in the Web interface.

### 4.3 Editing

Some of the tabs in the Explorer interface have an "edit" button. This allows registered users to manually correct certain types of annotations. Currently, the user can edit the tokenization view and on the derivation view. Clicking "edit" in the tokenization view gives an annotator the possibility to add and remove token and sentence boundaries in a simple and intuitive way, as Figure 4 illustrates. This editing is done in real-time, following the WYSIWYG strategy, with tokens separated by spaces and sentences separated by new lines. In the derivation view, the annotator can change part-of-speech tags and named entity tags by selecting a tag from a drop-down list (Figure 5).

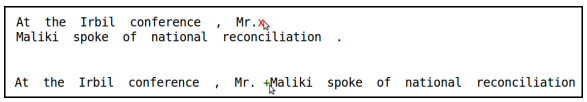


Figure 4: Tokenization edit mode. Clicking on the red '×' removes a sentence boundary after the token; clicking on the green '+' adds a sentence boundary.

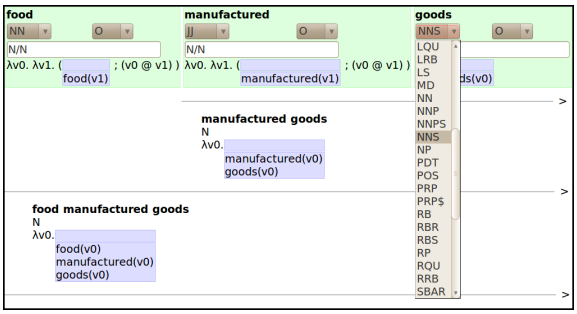


Figure 5: Tag edit mode, showing derivation with partial DRSs and illustrating how to adjust a POS tag.

As the updating daemon is running continually, the document is immediately reprocessed after editing so that the user can directly view the new annotation with his BOW taken into account. Re-analyzing a document typically takes a few seconds, although for very large documents it can take longer. It is also possible to directly rerun the NLP toolchain on a specific document via the "reprocess" button, in order to apply the most recent version of the software components involved. The GMB Explorer shows a timestamp of the last processing for each document.

We are currently working on developing new editing options, which allow users to change different aspects of the semantic representation, such as word senses, thematic roles, co-reference and scope.

## 5 Demo

In the demo session we show the functionality of the various features in the Web-based user interface of the GMB Explorer, which is available online via: `http://gmb.let.rug.nl`.

We show (i) how to navigate and search through all the documents, including the refinement of search on the basis of the lexical category or part of speech, (ii) the operation of the different view options, including the raw, tokenized, derivation and semantics view of each document, and (iii) how adjustments to annotations can be realised in the Web interface. More concretely, we demonstrate how boundaries of tokens and sentences can be adapted, and how different types of tags can be changed (and how that affects the syntactic, semantic and discourse analysis).

In sum, the demo illustrates innovation in the way changes are made and how they improve the linguistic analysis in real-time. Because it is a web-based platform, it paves the way for a collaborative annotation effort. Currently it is actively in use as a tool to create a large semantically annotated corpus for English texts: the Groningen Meaning Bank.

## References


Guillaume Artignan, Mountaz Hascoët, and Mathieu Lafourcade. 2009. Multiscale visual analysis of lexical networks. In *13th International Conference on Information Visualisation*, pages 685–690, Barcelona, Spain.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.

Johan Bos. 2008. Wide-Coverage Semantic Analysis with Boxer. In J. Bos and R. Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 277–286. College Publications.

J. Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann. 2003. The NITE XML toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, 35(3):353–363.

Central Intelligence Agency. 2006. *The CIA World Factbook*. Potomac Books.

John Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2008. Addressing the Resource Bottleneck to Create Large-Scale Annotated Texts. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 375–380. College Publications.

James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic.

Mike Dowman, Valentin Tablan, Hamish Cunningham, and Borislav Popov. 2005. Web-assisted annotation, semantic indexing and search of television and radio news. In *Proceedings of the 14th International World Wide Web Conference*, pages 225–234, Chiba, Japan.

U. Hahn, E. Buyko, K. Tomanek, S. Piao, J. McNaught, Y. Tsuruoka, and S. Ananiadou. 2007. An annotation type system for a data-driven NLP pipeline. In *Proceedings of the Linguistic Annotation Workshop*, pages 33–40, Prague, Czech Republic, June. Association for Computational Linguistics.

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: a community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 68–73, Stroudsburg, PA, USA.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT*. Kluwer, Dordrecht.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *In Proceedings of LREC 2004*, pages 2237–2240.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of english. *Journal of Natural Language Engineering*, 7(3):207–223.

Mark Steedman. 2001. *The Syntactic Process*. The MIT Press.