

The Groningen Meaning Bank

Johan Bos, Valerio Basile, Kilian Evang, Noortje Venhuizen, and Johannes Bjerva

Abstract The goal of the Groningen Meaning Bank (GMB) is to obtain a large corpus of English texts annotated with formal meaning representations. Since manually annotating a comprehensive corpus with deep semantic representations is a hard and time-consuming task, we employ a sophisticated bootstrapping approach. This method employs existing language technology tools (for segmentation, part-of-speech tagging, named entity tagging, animacy labelling, syntactic parsing, and semantic processing) to get a reasonable approximation of the target annotations as a starting point. The machine-generated annotations are then refined by information obtained from both expert linguists (using a wiki-like platform) and crowd-sourcing methods (in the form of a ‘Game with a Purpose’) which help us in deciding how to resolve syntactic and semantic ambiguities. The result is a semantic resource that integrates various linguistic phenomena, including predicate-argument structure, scope, tense, thematic roles, rhetorical relations and presuppositions. The semantic formalism that brings all levels of annotation together in one meaning representation is Discourse Representation Theory, which supports meaning representations that can be translated to first-order logic. In contrast to ordinary treebanks, the units of annotation in the GMB are texts, rather than isolated sentences. The current version of the GMB contains more than 10,000 public domain texts aligned with Discourse Representation Structures, and is freely available for research purposes.

Johan Bos
University of Groningen e-mail: johan.bos@rug.nl

Valerio Basile
University of Groningen e-mail: v.basile@rug.nl

Kilian Evang
University of Groningen e-mail: k.evang@rug.nl

Noortje J. Venhuizen
University of Groningen e-mail: n.j.venhuizen@rug.nl

Johannes Bjerva
University of Groningen e-mail: j.bjerva@rug.nl

1 Introduction

Data-driven approaches in computational semantics are still rare compared to approaches currently employed in syntactic parsing, where statistical methods dominate. This is not that surprising, given the fact that there are not many large annotated resources at our disposal that provide empirical information about various levels of semantic analysis, such as: anaphora, presupposition, scope, events, tense, thematic roles, animacy, named entities, word senses, ellipsis, discourse segmentation and rhetorical relations. It is challenging and time-consuming to create such annotated resources from scratch, and even more challenging to do so for multiple linguistic phenomena using a single semantic formalism.

Nonetheless, various semantically annotated corpora of reasonable size exist nowadays, including PropBank [50], FrameNet [3], and the Penn Discourse TreeBank [54]. However, efforts that combine various levels of annotation into one formalism are rare. One example is OntoNotes [34], a resource comprising syntax (in the style of the Penn Treebank, PTB [43]), predicate-argument structure (based on PropBank), word senses, and co-reference. Yet, what all these annotated corpora have in common is that they lack a level of formal meaning representation that combines various layers of semantic annotation. We believe, however, that a solid backbone of formal representation is important for driving semantic annotation, both from a theoretical point of view (for instance, for maintaining clarity and consistency), as well as from a practical point of view (e.g., enabling logical inference).

In this chapter, we describe the results of an ongoing effort to fill this gap: the Groningen Meaning Bank (GMB) project [6]. The aim of this project is to provide a large collection of semantically annotated English texts with deep semantic representations. One of its key objectives is to integrate phenomena into a *unified* formalism, instead of covering single phenomena in a linguistically isolated way. This will, we believe, provide a better handle on explaining dependencies between various ambiguous linguistic phenomena. Another key objective is to annotate *texts*, instead of isolated sentences—as is standard in existing treebanks, such as the PTB [43]. This allows us to deal with, for example, ambiguities on the sentence level that require the discourse context to be resolved. More specifically, the questions that drive our research on semantic annotation are:

1. What is a useful meaning representation and how can it interact with other linguistic levels of annotation?
2. To what extent can existing natural language processing software be used to create an annotated corpus?
3. How to obtain qualitatively good human annotations from multiple sources (experts and laymen)?
4. How can we ensure the largest distribution possible for the purpose of fostering scientific research?

We will answer these questions in this chapter in the following way. In Section 2 we motivate our choice of semantic formalism, which is rooted in Discourse Representation Theory [37]. Next, we introduce our annotation scheme for formal

meaning representations, including segmentation of words and sentences, and a description of all layers of linguistic information required to produce meaning representations in a systematic way (Section 3). Then, in Section 4, we outline our annotation method, which we dub *human-aided machine annotation*. We illustrate the toolchain of language technology components that we employ, and show how we apply information provided (mostly) by human annotators, who correct choices made by automated annotation methods. In this section we will also motivate our choice of data and explain how we manage it. In Section 5 we present two methods for acquiring annotations, obtained from two main sources of annotators: linguists, and non-experts. For the expert linguists, we have developed a wiki-like platform from scratch, because existing annotation systems (e.g., GATE [26], NITE [18], or UIMA [31]) do not offer the functionalities required for deep semantic annotation. For the non-experts, we introduce a crowd-sourcing method based on gamification. Finally, in Section 6, we take stock of what we have achieved so far, discuss current and potential applications, and provide a brief outlook into the future of meaning banking.

2 The Semantic Formalism: Discourse Representation Theory

Formal approaches to semantics have long been restricted to small or medium-sized fragments of natural language grammars. In the Groningen Meaning Bank, we aim to automatically deduce the meaning of large amounts of real-world texts. In this section we motivate our choice of formalism, grounded in Kamp’s Discourse Representation Theory, and show that it is a good candidate for combining high linguistic coverage with practical issues such as readability and the applicability for automated reasoning.

2.1 Background and Motivation

Discourse Representation Theory (DRT) is a widely applied dynamic theory of meaning representation that has been developed to provide a framework to include various linguistic phenomena, including the interpretation of discourse anaphora, temporal expressions and plural entities [36, 37]. The basic meaning-carrying units in DRT are Discourse Representation Structures (DRSs), which are recursive formal meaning structures that have a model-theoretic interpretation. This interpretation can be given directly [37] or via a translation into first-order logic [47]. This property is not only interesting from a theoretical point of view, but also from a practical perspective, because it permits the use of efficient existing inference engines (e.g. theorem provers and model builders) developed by the automated deduction community.

As the goal of the Groningen Meaning Bank is to provide deep semantic annotations, DRT is particularly suitable because it can be easily extended to incorporate a wide range of semantic phenomena. For the purpose of the GMB, we use a variant of DRT that uses a neo-Davidsonian analysis of events (via the VerbNet inventory of thematic roles [38]), accounts for presupposition projection in a revised version of van der Sandt’s [61] treatment of projection, using the Projective DRT framework [64], and incorporates rhetorical relations based on Segmented DRT [1, 2]. Let us have a closer look at these three extensions of the theory.

In a neo-Davidsonian take on event semantics, events are first-order entities characterised by one-place predicate symbols. Events are combined with their semantic arguments using an inventory of thematic roles, which are encoded as two-place relations between the event and its sub-categorised arguments or modifiers. We choose this way of representing events because it yields a more consistent analysis of event structure. We use VerbNet [38] as our inventory of thematic roles, because it contains a relatively small and well-defined set of roles. A few examples of VerbNet’s thematic roles are: *Agent* (a human or animate subject), *Experiencer* (a participant that is aware or experiencing something), and *Theme* (a participant in a location or undergoing a change of location).

Projective Discourse Representation Theory (PDRT) is an extension of DRT specifically developed to account for the interpretation of presuppositions and other projection phenomena, such as Potts’ [53] conventional implicatures [64, 65]. This formalism applies van der Sandt’s [61] idea of ‘presupposition projection as anaphora resolution’ by introducing projection variables (*labels* and *pointers*) that indicate the interpretation site of semantic content. In PDRT, each basic structure introduces a label, which can be bound by the pointers associated with the referents and conditions; the pointer of asserted content will be bound by its local context, while the pointer of projected content is either bound by an accessible context or occurs free. This way, no semantic content needs to be moved at the representational level, which aids incremental construction and increases the correspondence between the linguistic surface form and the representation of its meaning. The latter feature enhances the readability of the meaning representations, and hence facilitates semantic annotation.

In order to account for the rhetorical structure of texts, we use the widely applied DRT extension known as Segmented Discourse Representation Theory (SDRT), which aims at formalising the dynamic semantics of rhetorical relations [1, 2]. The variant of SDRT used in the Groningen Meaning Bank links discourse segments (i.e., DRSs) to each other via binary relations, resulting in a recursive structure that may again be embedded. The relations of SDRT can be divided into horizontal (*coordinating*) relations and vertical (*subordinating*) relations, reflecting different characteristics of textual coherence, such as the temporal order and the communicative intentions. The coordinating relations currently used in the GMB are: *continuation*, *narration*, *background* and *result*. The subordination relations are: *elaboration*, *instance*, *topic*, *explanation*, *precondition*, *commentary* and *correction*.

2.2 Dynamic Meaning Representations

One of the main principles of Discourse Representation Theory is that a DRS can play both the role of semantic content, and the role of discourse context [62]. The content of a DRS provides the precise model-theoretic meaning of a natural language expression, and the context it sets up aids in the interpretation of subsequent anaphoric expressions occurring in the discourse. This dynamic view on meaning results in a formalism that can be divided into three major components. The central component is a formal language defining Discourse Representation Structures (DRSs), the meaning representations for texts. The second component deals with the semantic interpretation of DRSs. The third component constitutes an algorithm that systematically maps natural language text into DRSs: the syntax-semantics interface. Below we describe how these components are formalised in our version of DRT.

2.2.1 Discourse Representation Structures

The syntax of DRSs is based on a type system. The basic semantic types in our inventory are e (individuals) and t (truth value). The set of all types is recursively defined in the usual way: if τ_1 and τ_2 are types, then so is $\langle \tau_1, \tau_2 \rangle$, and nothing except the basic types or what can be constructed via this recursive rule are types. Expressions of type e are either discourse referents, or variables. Expressions of type t are either basic DRSs, Segmented DRSs, or Combinatory DRSs.

$$\langle \text{exp}_e \rangle ::= \langle \text{ref} \rangle \mid \langle \text{var}_e \rangle$$

$$\langle \text{exp}_t \rangle ::= \langle \text{drs} \rangle \mid \langle \text{sdrs} \rangle \mid \langle \text{cdrs} \rangle$$

Basic DRSs ($\langle \text{drs} \rangle$) consist of a set of referents and a set of conditions. In addition, the basic DRSs are decorated with the projection variables from PDRT ($\langle \text{pvar} \rangle$), i.e., each DRS is associated with a *label*, and each referent and condition obtains a *pointer* that can be bound by a DRS label to indicate the interpretation site of the semantic content (following [64]). Segmented DRSs ($\langle \text{sdrs} \rangle$) are recursive structures that combine two expressions of type t by means of coordinating or subordinating relations. Combinatory DRSs ($\langle \text{cdrs} \rangle$) combine type t expressions using one of the merge operators (assertive merge (+) or projective merge (\times); see [64]) or apply function application (@), which turns a complex type into a type t expression [47]. Following the conventions in the DRT literature, we will visualise DRSs in their usual box-like format.

$$\langle \text{drs} \rangle ::= \langle \text{pvar} \rangle : \frac{(\langle \text{pvar} \rangle, \langle \text{ref} \rangle)^*}{(\langle \text{pvar} \rangle, \langle \text{condition} \rangle)^*}$$

$$\langle \text{sdrs} \rangle ::= \frac{k_1 : \langle \text{exp}_t \rangle \quad k_2 : \langle \text{exp}_t \rangle}{\text{coo}(k_1, k_2)} \mid \frac{k_1 : \langle \text{exp}_t \rangle \quad k_2 : \langle \text{exp}_t \rangle}{\text{sub}(k_1, k_2)}$$

$$\langle \text{cdrs} \rangle ::= (\langle \text{exp}_t \rangle + \langle \text{exp}_t \rangle) \mid (\langle \text{exp}_t \rangle \times \langle \text{exp}_t \rangle) \mid (\langle \text{exp}_{(\alpha, t)} \rangle @ \langle \text{exp}_\alpha \rangle)$$

The discourse referents in a DRS ($\langle \text{ref} \rangle$) can be seen as a record of topics mentioned in a sentence or text. The conditions ($\langle \text{condition} \rangle$), in turn, tell us how the discourse referents relate to each other, and put further semantic constraints on their interpretation. We distinguish between basic and complex conditions. The basic conditions express properties of discourse referents or relations between them:

$$\begin{aligned} \langle \text{condition} \rangle & ::= \langle \text{basic} \rangle \mid \langle \text{complex} \rangle \\ \langle \text{basic} \rangle & ::= \langle \text{sym}_1 \rangle (\langle \text{exp}_e \rangle) \\ & \quad \mid \langle \text{sym}_2 \rangle (\langle \text{exp}_e \rangle, \langle \text{exp}_e \rangle) \\ & \quad \mid \langle \text{exp}_e \rangle = \langle \text{exp}_e \rangle \\ & \quad \mid \text{card}(\langle \text{exp}_e \rangle) = \langle \text{num} \rangle \\ & \quad \mid \text{timex}(\langle \text{exp}_e \rangle, \langle \text{sym}_0 \rangle) \\ & \quad \mid \text{named}(\langle \text{exp}_e \rangle, \langle \text{sym}_0 \rangle, \text{class}) \end{aligned}$$

Here $\langle \text{sym}_n \rangle$ denotes an n -place predicate symbol, and $\langle \text{num} \rangle$ a cardinal number. Nouns, verbs, adverbs and adjectives introduce a one-place relation; prepositions and thematic roles introduce two-place relations. Since our DRS-language uses a neo-Davidsonian event-structure, there are no ternary or higher-place relations. The cardinality condition is used for numerals, the timex condition for temporal entities, and the naming condition for proper names of a certain class. The equality condition explicitly states that two discourse referents denote the same individual.

Next we turn to complex conditions. For convenience, we split them into unary and binary complex conditions. The unary complex conditions have one DRS as argument and represent negation, the modal operators expressing necessity and possibility, and a ‘‘hybrid’’ condition representing a propositional DRS [12]. The binary conditions have two DRSs as arguments and represent the conditions for implication, disjunction, and interrogative constructions:

$$\begin{aligned} \langle \text{complex} \rangle & ::= \langle \text{unary} \rangle \mid \langle \text{binary} \rangle \\ \langle \text{unary} \rangle & ::= \neg \langle \text{exp}_r \rangle \mid \Box \langle \text{exp}_r \rangle \mid \Diamond \langle \text{exp}_r \rangle \mid \langle \text{ref} \rangle : \langle \text{exp}_r \rangle \\ \langle \text{binary} \rangle & ::= \langle \text{exp}_r \rangle \Rightarrow \langle \text{exp}_r \rangle \mid \langle \text{exp}_r \rangle \vee \langle \text{exp}_r \rangle \mid \langle \text{exp}_r \rangle ? \langle \text{exp}_r \rangle \end{aligned}$$

The unary complex conditions are mostly activated by negation particles or modal adverbs. The hybrid condition is used for the interpretation of verbs expressing propositional content and other linguistic phenomena that take sentential complements. The binary complex conditions are triggered by conditional statements, certain determiners, and questions.

Finally, we turn to expressions with complex types. As described above, Combinatory DRSs ($\langle \text{cdrs} \rangle$) may apply function application to complex types, in order to obtain DRSs of type t [47]. There are three kinds of expressions with complex types: variables ranging over complex types, λ -abstraction, and function application. For function application, we follow the notational convention introduced by Blackburn & Bos [9], using ‘@’ instead of brackets:

$$\langle \text{exp}_{\langle \alpha, \beta \rangle} \rangle ::= \langle \text{var}_{\langle \alpha, \beta \rangle} \rangle \mid \lambda \langle \text{var}_\alpha \rangle . \langle \text{exp}_\beta \rangle \mid (\langle \text{exp}_{\langle \gamma, \langle \alpha, \beta \rangle} \rangle) @ \langle \text{exp}_\gamma \rangle$$

In the GMB, complex types are used to represent the semantics of sub-sentential expressions, such as words or combinations of words, which still need to be combined

with other expressions in order to obtain a complete DRS representation of type t . An example of how these complex type DRSs are represented in the GMB is shown in Figure 1 below.

2.2.2 Semantic Interpretation

The semantic interpretation of the meaning representations in the GMB is carried out by translating DRSs into formulas of first-order logic. The DRS-language employed in our large-scale lexicon is very similar to the language formulated by Kamp and Reyle [37], but differs on some crucial points. On the one hand, it is more restrictive, as it leaves out the so-called “duplex conditions” that Kamp and Reyle employ for representing quantifiers like “most”, because these conditions do not all permit a translation to first-order logic. On the other hand, our DRS-language forms an extension to Kamp and Reyle’s, as it includes a number of modal operators on DRSs, including ones that are employed to analyse sentential complements. Moreover, our version of DRSs includes the projection variables from PDRT. As we have shown, however, these structures can be directly translated into standard DRSs without projection variables [64]. The resulting DRS-language is known to have a translation to ordinary first-order formulas. Examples of such translations are given in [37, 47], and [10], disregarding the modal operators. A translation incorporating the modal operators is given by [13]. We will not give the translation in all its detail here, but interested readers are referred to the articles cited above.

2.2.3 The Syntax-Semantics Interface

As a preliminary to a compositional semantics, we need syntactic structure of some kind. The syntax-semantics interface employed in the GMB is based on categorial grammar. More precisely, we use a broad-coverage version of Combinatory Categorical Grammar, or CCG for short [58]. A categorial grammar lends itself extremely well to the task of specifying a compositional semantics because CCG is lexically driven and has only few “grammar” rules; the combinatory rules include forward ($>$) and backward application ($<$), composition (B), and type raising (T).

CCG’s type-transparency principle, which states that each syntactic category corresponds to a unique semantic type, is helpful in ensuring that the outcome of a derivation always corresponds to an expression of the desired semantic type. The syntactic categories used in the GMB are based on the ones introduced in CCG-bank [33], a large collection of CCG derivations for a newswire corpus. There are two types of syntactic categories: base categories and functor categories. The base categories employed in the GMB are S (sentence), NP (noun phrase), N (noun) and PP (prepositional phrase). The S and NP categories correspond to DRS expressions of type $\langle\langle e, t \rangle, t\rangle$, and the N and PP categories correspond to expressions of type $\langle e, t \rangle$. Functor categories are composed out of the base categories with the forward and backward slash, where the direction of the slash indicates whether the

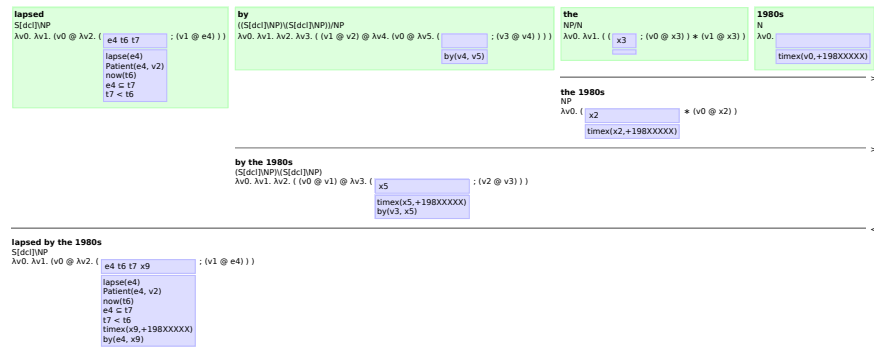


Figure 1 CCG derivation, decorated with lexical semantics in the form of λ -DRSs.

argument is to its left or its right. For instance, a verb phrase is represented with functor category $S \backslash NP$, which requires a noun phrase on its left and has semantic type $\langle \langle \langle e, t \rangle, t \rangle, \langle \langle e, t \rangle, t \rangle \rangle$. An adjective, on the other hand, can be represented with the functor category N/N , which requires an expression of category N on its right, and has $\langle \langle e, t \rangle, \langle e, t \rangle \rangle$ as the corresponding semantic type.

Because most of the work in a lexicalised grammar such as CCG is done in the lexicon, syntactic annotation can be carried out almost exclusively on the word level. This makes it a convenient framework to use in the context of developing the GMB, because there is no need to annotate syntactic structures. Furthermore, the availability of robust parsers trained on CCGbank [21] make CCG a practically motivated choice. Figure 1 shows a CCG derivation as produced by the GMB, together with the associated semantic representations.

2.3 Meaning Representations for Real-World Texts

Above we have described the representational semantic formalism used in the GMB, which originates from Discourse Representation Theory, and follows to a great extent the theory as formulated by its originator Hans Kamp. However, it deviates on certain points, as it comprises:

- a neo-Davidsonian view on representing event structures;
- projection pointers indicating the interpretation site of semantic content (following Projective DRT, see above);
- rhetorical relations between DRSs (following Segmented DRT, see above);
- a syntax-semantics interface based on categorial grammar (CCG) and type-theory.

With these ingredients, we can represent a wide range of semantic phenomena that may occur in texts, for instance: coreference, event structure, presupposition projection, rhetorical structure, ellipsis, and temporal organisation. As mentioned before,

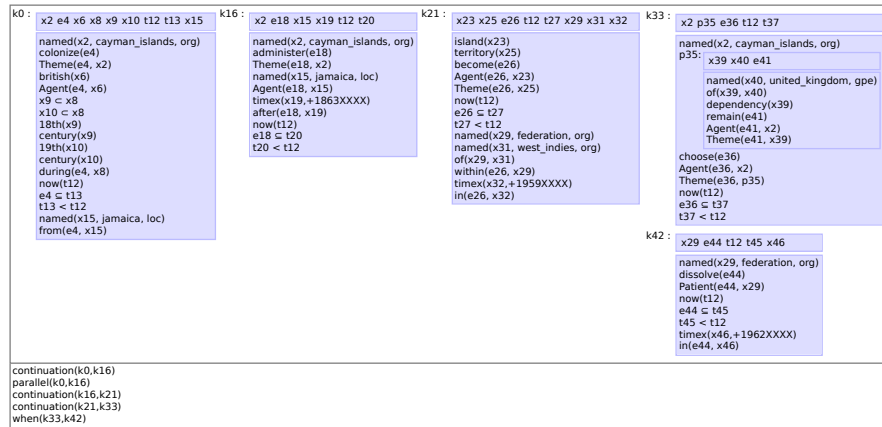


Figure 2 An example of a semantic representation of a text in the GMB, with DRSs representing discourse units.

one of the trademarks of the GMB is that it provides the information of various layers of meaning within a single representation format: a DRS. This is illustrated in Figure 2, which shows an example of the complex DRS representations, as they are currently shown in the Groningen Meaning Bank. Note that in the current version of the GMB, the projection variables from PDRT are only represented internally.

Importantly, the application of a theoretical formalism to real-world texts (instead of made-up sentences found in the semantic literature) will inevitably illustrate its shortcomings. Indeed, there are still several aspects of linguistic meaning that are currently hard to represent in a formal framework like DRT. For example, dialogues may introduce an additional semantic layer by means of quotation, which can not straightforwardly be represented in a DRS. Similarly, the embedded meaning of complex named entities, such as song titles, is currently not part of the meaning representation of the expression. In the future, we will aim to extend the semantic formalism so that it can account for such cases.

3 An Annotation Scheme for Meaning Banking

The GMB comprises several levels of annotation below the semantic analysis described in the previous section. We start annotating at the token level, segmenting the texts into separate words and sentences. At the word level, there are several layers of annotation, including lexical categories (part-of-speech and syntactic categories), classes of lexical meaning (named entities, animacy classes and word senses), coreference information, scope, thematic roles and implicit relations. In this section we will present each of these levels in detail, compare our approach with existing annotation schemes, and point out some difficulties encountered during annotation.

```

It didn't matter if the faces were male,
SIOIIIIIIOTIIIIIIOTIIOTIIIIOTIIIIOTIIIIIO
female or those of children. Eighty-
TIIIIIOIOTIIIIIOIOTIIOTIIIIIIITOSIIIIIO
three percent of people in the 30-to-34
IIIIIOIOTIIIIIIOTIIOTIIIIIIOTIIOTIIIIIIIO
year old age range gave correct responses.
TIIIOIIOIOTIIIIIOIOTIIIIOTIIIIIIOTIIIIIIIT

```

Figure 3 Example of IOB-labelled characters, with two kinds of B-tags: S for the beginning of a sentence, and T for the beginning of a token.

3.1 Segmentation of Raw Text into Words and Sentences

The final result of any annotation effort depends on its initial input — any mistakes occurring in the low-level segmentation of the data are often hard to recover in more fine-grained semantic annotation. We therefore designed the GMB with segmentation tools that are flexible and dynamic, in order to make sure that changing any segmentation decisions or conventions later on, does not affect the annotations carried out at other layers. Additionally, the method used in the GMB provides an alignment between raw and tokenised text, which makes mapping the tokenised version back onto the actual source unproblematic.

The process of segmentation divides the raw text into word tokens and sentence tokens. We use an IOB (“Inside–Outside–Beginning”) tagging scheme combining word and sentence segmentation decisions [27]. IOB tagging is widely used in tasks identifying chunks of tokens; we here apply it to identify chunks of characters. Characters *inside* tokens are labelled with ‘I’, and characters *outside* of tokens are labelled with ‘O’. For characters at the *beginning* of tokens, we use two different types of tags: ‘S’ at sentence boundaries, and ‘T’ to mark the beginning of a token. The main advantages of the scheme are that it can account for discontinuous tokens, e.g., hyphenated words at line breaks, and that it is possible to mark the beginning of a new token in the middle of a typographic word, as is the case, e.g., in *did|n’t*. An example is given in Figure 3 (from [27]).

The segmentation of text into words and sentences is generally a straightforward task, but there are some notorious hard cases. Below we describe some of these cases, and how they are resolved in the GMB, which in most cases follows the segmentation conventions as used in the PTB [43].

- **Unclear sentence boundaries.** Sentence boundaries are in general clearly marked by means of punctuation symbols. However, in some cases the punctuation is missing, due to errors or simply because they are not required, as is the case for section headers, for instance, or if a sentence ends with an abbreviation that contains a haplographical full stop. In the GMB, these are all permissible sentence boundaries. In the case of quoted contexts, multiple sentences can be part of a single quote; in the GMB, we choose to split these into separate sentence tokens.

- **Hyphenated compound expressions.** Examples of this kind are adjectives, such as ‘fifty-eight-year-old’, ‘colder-than-normal’ and ‘Blizzard-like’, past participles, such as ‘British-controlled’ and ‘Cypriot-occupied’, and split constituents in coordinated constructions, such as ‘short- and medium-term loans’. Currently, these are not split into separate tokens in the GMB, because of the difficulties that would arise in later stages of the NLP toolchain. Nevertheless, some of the examples above are systematic and have a compositional semantics, and therefore remain a challenge for future research.
- **Non-hyphenated compound expressions.** In some cases, even two separate words can be interpreted as constituting a single token. This is the case, for example, in expressions like ‘New York’ and ‘San Francisco’, where there does not seem any compositional semantics at play in determining the referent. Currently, we follow the convention used in the PTB, where these expressions are treated as two separate tokens. It is interesting to note, however, that our annotation scheme is flexible and would allow for this type of expression to be annotated as a single token (i.e., by tagging the intermediate whitespace as ‘O’).

3.2 *Annotating Lexical Categories*

The meaning representations of the GMB are constructed following the compositionality principle. This means that a lot of information that is required for disambiguation is coming from the individual word tokens. For specifying the lexical annotations, we use the token-ID—a number ($1000 * m + n$) identifying the n -th token in the m -th sentence (this method assumes that the number of words in a sentence does not exceed a thousand). Internally, tokens are identified by pairs of character offsets, so that lexical information is not lost when the tokenisation changes. The first level of lexical information is constituted by the lexical categories, which include Part-of-Speech tagging and the syntactic categories from CCG.

3.2.1 **Part-of-Speech tagging**

The first level of lexical annotation is assigning each word token a syntactic category through Part-of-Speech tagging. The Part-of-Speech (POS) categories form an important source of information for later syntactic and semantic processing. The GMB employs the tagset and most of the conventions introduced by the PTB [43], which was later adopted (and extended) by CCGbank [33]. In the GMB we aim to adhere as closely as possible to the original tagset introduced during the development of the PTB. Nonetheless, the result of the annotation process shows various inconsistencies due to different reasons [42]. These result, for example, from unclear annotation guidelines for certain linguistic phenomena, hard linguistic cases, or inconsistencies in the statistical model of the POS-tagger that is used in the GMB. Notorious examples here are the distinction between participles and adjectives, tag-

ging of time expressions, and distinguishing past-tense verbs from past participles. Fortunately, however, most of these inconsistencies do not have a direct impact on the semantic representations.

An interesting case in point are complex named entities, such as *Secretary of State Condoleezza Rice*, or *President of The Thai Rice Exporters Association*. In these examples, it is debatable whether the prepositions and articles that are part of the name ought to be tagged with their basic tag (IN and DT, respectively), or as proper names (NNP). Similarly, it is unclear whether capitalised nouns such as ‘State’ and ‘President’ should be tagged as normal nouns (NN) or proper names (NNP). This issue is even more pressing for titles of songs or other works of art, as in the sentence *Fats Domino wrote a song called The Fat Man*. Here, the expression “The Fat Man” seems to make a contribution both as the name of the song (which would correspond to the POS sequence: NNP, NNP, NNP), and with its literal meaning (which, in turn, corresponds to the POS sequence: DT, JJ, NN). Ideally, both of these POS representations would be reflected in the annotation. However, the GMB only applies one layer of POS information, so all complex named entities are annotated as a sequence of NNP-tags. For these hard cases, the GMB follows the convention used in the PTB, but it is clear that POS annotation for embedded expressions is an interesting issue for future research in semantic annotation.

3.2.2 Syntactic categories

Based on the POS tags, we can assign each word a syntactic category from the syntactic formalism used in the GMB: Combinatory Categorical Grammar. As described in Section 2.2.3, CCG is a lexicalised framework where syntactic categories are composed out of a few base categories (S, NP, N, PP), and slashes of functor categories indicate the direction of arguments. In addition, the S category is decorated with a feature indicating sentence mood, or aspectual status, following the conventions from CCGbank [33].

The choice of CCG is not accidental nor arbitrary. CCG supports a consistent compositional semantics because of its type-transparency principle. This entails that every basic syntactic category is mapped to precisely one semantic type. The semantic types of functor categories can be computed recursively from the core semantic types of the base categories. The combinatory rules in CCG have a fixed semantic interpretation. These two properties of CCG together form a very convenient and systematic platform for meaning banking on a large scale.

3.3 Annotating Lexical Meaning

There are three different types of lexical meaning that are included in the GMB. These are named entity categories, animacy properties, and word senses. In this section we will look at these three layers of meaning annotation in more detail.

3.3.1 Named Entity Types

Types of named entities are important to semantically distinguish locations from persons, artefacts from organisations, and so on. Given the way we aim to obtain our annotations, we have opted for an annotation scheme that is both simple and rudimentary. The set of named entity types used in the GMB is partly based on Sekines Extended Named Entities [57], with some modifications. First of all, we only use a subset of the types. In particular, we leave out the fine-grained types, resulting in a more coarse-grained scheme. Currently, the annotation schema for named entities adheres to the following conventions:

- Nested named entities are not tagged separately. Rather, we tag only the “outer” NE tag (the head), e.g., all tokens in the expression “Los Angeles Lakers” are tagged as ‘Organisation’;
- Honorifics (i.e., titles etc.) are POS-tagged as nouns (NN), which means that they are not considered names and are thus not tagged as NEs;
- Time and numerical expressions are annotated on separate layers;
- The GPE (Geo-Political Entity) tag is used to label expressions that can be interpreted both as a location and an organization;
- A named entity serving as a pre-modifier, for example *Korean*, is POS-tagged as adjectives (JJ), but also obtains a named entity category (in this case, GPE).

As Table 1 illustrates, we adopt a total of seven named entity types. Named entities are represented in the semantic representations by the *named* condition, of which one argument is reserved to indicate the type of named entity. The meaning of a DRS condition $named(X, \text{“John”}, PER)$ can be paraphrased as: X is a person, and X is named “John”. In general, proper names are used to select particular individuals. In some case, however, names refer to classes, as in the case of ‘CRJ-200’ in Table 1 above, which refers to a specific *class* of planes. Similarly, in the sentence *Sammy is a parrot, a Hyacinth Macaw*, the expression *Sammy* is a proper name (referring to a particular individual), and *Hyacinth Macaw* a class name (referring to a particular species of parrot). In the GMB, we currently do not make a distinction between class names and proper names (*Hyacinth Macaw* in the example above will receive the NE-tag ‘NAT’).

3.3.2 Animacy

Animacy is a semantic property of nouns, which denotes whether (or to what extent) the referent of that noun is alive, human-like or even cognitively sophisticated. Even though animacy is rarely overtly marked in English, it influences the choice of various grammatical structures, including dative alternation [16], genitive constructions [59], and active and passive voice [56]. Moreover, it has been shown that animacy plays an important role in anaphora resolution [49, 41] and verb argument disambiguation [25].

Table 1 Named Entity tagset used in the GMB, illustrated with examples.

| Tag | Description | Examples |
|-----|----------------------|---|
| PER | Person | Mr. Putin 's talks in Egypt made him the first Russian leader to ... |
| GEO | Location | Mr. Putin 's talks in Egypt made him the first Russian leader to ... |
| ORG | Organisation | Google will present its annual report on Saturday. |
| TIM | Time | Google will present its annual report on Saturday . |
| EVE | Event | Hurricane Katrina slammed into southeast Florida Thursday. |
| ART | Artefact | The plane was a Canadian-made CRJ-200 . |
| NAT | Natural | The deadly H5N1 strain was found in a dead bird. |
| GPE | Geo-Political Entity | Mr. Putin's talks in Egypt made him the first Russian leader to ... |

Table 2 Animacy tagset used in the GMB, based on [67].

| Tag | Description | Examples |
|-----|--------------|--|
| HUM | Human | Mr. Calderon said Mexico has become a worldwide leader ... |
| ORG | Organisation | Mr. Calderon said Mexico has become a worldwide leader ... |
| ANI | Animal | There are only about 1,600 pandas still living in the wild in China. |
| LOC | Place | There are only about 1,600 pandas still living in the wild in China . |
| NCN | Non-concrete | There are only about 1,600 pandas still living in the wild in China. |
| CNC | Concrete | The wind blew so much dust around the field today. |
| TIM | Time | The wind blew so much dust around the field today . |
| MAC | Machine | The astronauts attached the robot , called Dextre, to the ... |
| VEH | Vehicle | Troops fired on the two civilians riding a motorcycle ... |

The tagset for animacy used in the GMB is based on the one proposed by Zaenen et al. [67], who present an annotation scheme for animacy consisting of nine categories, with a few additional tags for cases in which annotators were uncertain (see Table 2). This scheme can be arranged hierarchically, so that the classes ‘Concrete’, ‘Non-concrete’, ‘Place’ and ‘Time’ are grouped as inanimate, while the remaining classes are grouped as animate. With the exception of the additional tags for uncertain cases, this is the tag set used in the GMB. We assign animacy tags to all nouns and pronouns. Similarly to our tagging convention for named entities, we assign the same tag to the whole NP, so that *wagon driver* is tagged with HUM, although *wagon* in isolation would be tagged with CNC.

Problematic cases occur when, e.g., animals behave in a human-like manner. This happens rather frequently in one of our subcorpora (“Aesop’s fables”), where, for example, lions and hares are conversing with each other. In such cases, if clear human-like behaviour is exhibited, we have opted to tag these animals as HUM. This corresponds to the tagging guidelines used in [67].

3.3.3 Word Senses

In the GMB, tokens that are POS-tagged as either noun, verb, adjective or adverb are also associated with a word sense tag. Word senses are expressed as WordNet 3.1

synset identifiers [29]. In order to get an improvement over the coverage provided by WordNet, we plan to extend the layer of word sense annotation with links to DBPedia¹, possibly by exploiting the alignment provided by the UBY resource [30]. The use of WordNet synsets facilitates the development of a multilingual GMB, where links at the word level to languages other than English are provided by cross-lingual alignment resources such as MultiWordNet [52] or BabelNet [48]; see also the discussion in Section 6.3.

3.4 Annotating Contextual Meaning

In this section we explain how non-lexical meanings are annotated in the GMB. We will have a closer look at how co-reference information, thematic roles, and quantifier scope is embedded in the GMB framework.

3.4.1 Co-reference Information

Two or more noun phrases that denote the same entity are considered to be *co-referential*. In the GMB co-reference is annotated at the word token level, as a relation between the target word and the word that it co-refers with (the *antecedent*). Each referential expression that has a co-referential antecedent is annotated with the token-ID of the antecedent. In some cases, multiple correct antecedents are available; we call this a *co-reference chain*. Since the semantic analyser treats co-reference as a transitive property, the co-reference chains will be recognised and treated as introducing a single entity. Similarly, multi-word referential expressions, such as “President Barack Obama”, are treated as introducing a single entity at the semantic level. Therefore, each word that is part of a multi-word expression can serve as an antecedent for co-reference, or introduce a co-reference relation itself; at the semantic level this information will be extended to the entire multi-word expression. As a rule of thumb, however, we identify the antecedent that is closest to the referential expression as the correct antecedent.

Currently, pronouns, definite noun phrases, and proper names are being annotated with co-reference information in the GMB. Pronouns are the most paradigmatic cases of co-referential expressions and hardly occur in a context in which they do not have an antecedent. Definite noun phrases and proper names, on the other hand, often occur without any antecedent, in which case they are annotated with the co-reference tag ‘null’ (indicating that no antecedent can be selected). Definite descriptions also differ from proper names and pronouns (with the exception of the neutral pronoun “it”) with respect to the possible antecedents that they allow; whereas proper names and pronouns generally refer to entities (introduced by

¹ <http://dbpedia.org/>

nouns), definite descriptions may also refer to more abstract entities introduced, for example, by verbs (e.g., “to meet”–“the meeting”).

The current method for annotating co-reference in the GMB does not specify the relation between the referential expression and its antecedent, nor does it allow for the annotation of multiple antecedents. This means that the current format does not allow for annotating split antecedents of expressions referring to plural entities (e.g., “Britney Spears and Kevin Federline” – “the couple”). Moreover, we do not explicitly annotate cases of *bridging*, where an expression relates to an antecedent via a relation that is not identity (e.g., “Iran”–“the government”, where the latter expression can be paraphrased as “the government of Iran”). We are currently investigating whether the distinction between co-reference and bridging can be derived automatically, either based on features from the expressions involved (e.g., number and animacy classification), or based on external resources such as WordNet.

3.4.2 Thematic Roles and Implicit Relations

Semantic relations are relations between two entities, of which one is the internal and one the external entity. In the GMB semantic relations are two-place relations between discourse referents. The internal entity is usually an event, triggered by a verb; the external entity is usually triggered by a noun phrase. External entities are realised by arguments or adjuncts—annotation of roles differs with respect to whether external entities are arguments or adjuncts. Semantic relations are encoded in various annotated corpora including PropBank [50], VerbNet [38], FrameNet [3] (at a more detailed level than VerbNet, but it has a more limited coverage), and NomBank [44], the latter providing semantic roles for nouns rather than verbs. In the GMB there are two kinds of semantic relations that are annotated explicitly: thematic roles (adopted from VerbNet [38]), and implicit relations (relations that are not overtly expressed in the text).

Thematic roles are annotated in the GMB using a lexicalised approach [15], again taking advantage of CCG as syntactic formalism. In CCG, verbs (and nouns) encode all their arguments inside their lexical category, which means that we can divide tokens into those that trigger (a finite, ordered set of) semantic roles and those that do not. Annotation then boils down to assigning the correct roles to each token that introduces them. The possible roles can be directly derived from VerbNet [38], and the number of roles for categories associated with verbs is determined by the number of arguments encoded in the CCG category. Hence, there is no need to explicitly select the entities that play a semantic role, because syntax will take care of that. This makes annotation of roles in the GMB not only easier than in other approaches, it also makes it more flexible, because one could even annotate correct roles for a clause whose syntactic analysis is incorrect. The approach is illustrated in Table 3.

The types of implicit relations that are annotated in the GMB are those occurring in noun-noun compounds, possessive constructions, and temporal modifiers. The inventory of relations is based on English prepositions. For instance, in *The Apple*

Table 3 Mapping VerbNet roles to CCG categories. Example taken from [15].

| Class | Sense | VerbNet frame | Enhanced CCG category |
|--------------|-------|---------------------------------|---------------------------------------|
| build-26.1 | 1 | Agent V | S\NP:agent |
| build-26.1 | 1 | Agent V Product | (S\NP:agent)/NP:product |
| build-26.1 | 1 | Material V Product | (S\NP:material)/NP:product |
| build-26.1-1 | 1 | Asset V Product | (S\NP:asset)/NP:product |
| build-26.1 | 1 | Agent V Product {from} Material | ((S\NP:agent)/PP:material)/NP:product |
| build-26.1-1 | 1 | Agent V Product {for} Asset | ((S\NP:agent)/PP:asset)/NP:product |
| base-97.1 | 8 | Agent V Theme {on} Source | ((S\NP:agent)/PP:source)/NP:theme |

spokesman announced Wednesday that its new products will be released this week, there are four implicit relations: (spokesman) **of** Apple, (announced) **on** Wednesday, (products) **by** Apple, (released) **in** this week. Annotating these relations is implemented as another layer on the word token level.

3.4.3 Scope Ambiguities

A correct treatment of scope-bearing operators such as quantifiers, modality and negation is crucial for constructing accurate meaning representations. A lot of work has been done to describe the scope alternation behaviour of quantifiers, in particular, and to construct underspecified meaning representations from which all (theoretically) possible readings of a sentence containing them can be enumerated [11, 22]. Since the goal of the GMB is to provide a single, fully specified meaning representation for each text, an underspecification mechanism is not required. Scope is instead specified by manual annotation via an additional layer of tags on categories that *mediate* scope interactions between their arguments, i.e. verbs and prepositions. A pilot study on the GMB [28] showed that clear deviations from the default scope order of verbal arguments (subject > objects in surface order) is very rare (only 12 of 206 cases). The annotation effort in the GMB has therefore been focused on scope interactions mediated by prepositions in combination with universally-quantifying determiners, as exemplified in Table 4.

We use two different scope tags for prepositions: *Inverting*, indicating that the modifier takes wide scope, and *Default*, indicating that the modified constituent takes wide scope. The rest of the work is done by the lexical semantics of prepositions, which is chosen according to the scope tag. It determines the scope order in which arguments end up in the final meaning representation [28].

In some cases it is not clear what scope order is expressed in a sentence. In such cases, we generally prefer annotations resulting in the logically weaker reading. For example, in “the International Banking Repeal Act of 2002 resulted in *the termination of all offshore banking licenses*”, we could either assume a separate termination

Table 4 Prepositions modifying NPs and VPs, mediating default and inverted quantifier scope.

| Modifying Scope | Example |
|-----------------|---|
| NP | Default <i>All such attacks by drone aircraft</i> are believed to be carried out by U.S. forces. [76/0357] |
| NP | Inverting Finally the gorgeous jewel of the order, gleaming upon <i>the breast of every member</i> , suggested “your Badgesty,” which was adopted, and the order became popularly known as the Kings of Catarrh. [72/0696] |
| VP | Default NATO says militants surrounded the outpost, <i>firing from all directions with rocket-propelled grenades, small arms and mortars</i> . [92/0311] |
| VP | Inverting Jobs <i>grew in every sector except manufacturing</i> , with much of the growth due to hurricane clean-up efforts in Florida. [97/0059] |

event for each banking license or a single termination event for them all; we prefer the former by giving the universal quantifier wide scope.

4 Constructing the Meaning Bank

The creation of a resource like the Groningen Meaning Bank involves several stages, including the collection of data for meaning annotation, the selection and development of NLP tools for automatically analysing the data, and choosing the right way to store and evaluate the annotations. In this section, we describe each of these stages in the development of the GMB.

4.1 Gathering Raw Linguistic Data

A primary aim of the Groningen Meaning Bank is to provide both training data and a testbed for the development of statistical algorithms for semantic analysis, in the same way that treebanks have played a crucial role in the development of robust syntactic parsers. In order to be useful for the development of statistical techniques, a resource should be sufficiently large and provide high-quality annotations. Of course, there is a trade-off involved here: the bigger a resource, the more costly it is to provide high-quality annotations. The release policy of the GMB aims to provide a useful trade-off: stable releases are frequent (so far 2–3 times a year), each one larger than the previous one and with a higher number of manual corrections than the last.

In the area of corpus linguistics, two central properties that are desired of a corpus are representativeness of linguistic data and balance [55]. In Natural Language Processing research, on the other hand, these properties are only of secondary importance, as the development of NLP techniques often focuses on a relatively limited range of text types, and generalisation is done in a step by step fashion in order to

Table 5 The size of the GMB, as of January 20, 2015

| Subcorpus | Genre | Documents | Sentences | Tokens | Sent./Doc. | Tok./Sent. |
|--------------------|-----------|-----------|-----------|-----------|------------|------------|
| Voice of America | newspaper | 9,207 | 57,147 | 1,238,678 | 6.2 | 21.7 |
| CIA World Factbook | almanac | 514 | 4,428 | 112,535 | 8.6 | 25.4 |
| Aesop's fables | fable | 224 | 948 | 23,106 | 4.2 | 24.4 |
| Jokes | humour | 122 | 442 | 7,537 | 3.6 | 17.1 |
| MASC | misc. | 35 | 291 | 6,991 | 8.3 | 24.0 |
| Total | | 10,102 | 63,256 | 1,388,847 | 6.3 | 22.0 |

prevent underspecification of linguistic phenomena. In the GMB, we therefore chose to limit the available text types to English running texts, excluding other genres such as transcribed speech or dialogue. Moreover, since unhampered availability of data is of great importance for enabling collaboration in research on a broad scale and for verifying research results, the GMB only includes documents that can be freely redistributed without charge or signing license agreements. In practice, this choice limits the selection to texts in the public-domain or distributed under permissive licenses. The current version of the GMB includes the following sub-corpora:

- *Voice of America*, a newspaper published by the US Federal Government (available at <http://www.voanews.com>);
- A selection of *MASC* documents of the Open American National Corpus [35];
- A collection of Aesop's fables (<http://www.aesopfables.com>);
- A number of humorous stories and jokes (<http://www.basicjokes.com>);
- A series of country descriptions from the *CIA World Factbook* [19], in particular the Background and Economy sections.

All of these texts have been cleaned (e.g. by removing HTML tags) by means of custom scripts, with the exception of the *MASC* articles, which are freely available for download from the project's website.² On occasion, data from new sources is added to the corpus. Current candidates for inclusion in the GMB are the EMEA 0.3 parallel corpus from the European Medicines Agency [60] (1,948 documents) and the first part (15 sections) of the British Nationality Act 1981³ which has been studied using the GMB toolchain of analysis [66].

In order to make sure that the documents in the GMB, along with their annotations, remain of a high quality, all newly collected documents are manually reviewed, and filtered based on appropriateness for semantic analysis (e.g., filtering out offensive or linguistically malformed texts). The current version of the GMB comprises about 10K accepted documents (see Table 5), containing 62K sentences and 1.3 million tokens.

Finally, a word about document identification in the GMB. The GMB is divided into a hundred parts. Each document is identified using 6 digits, of which the first

² <http://www.anc.org/data/masc/downloads/data-download/>

³ <http://www.legislation.gov.uk/ukpga/1981/61>

two indicate the part it belongs to, e.g. document 16/0690 is the 690th document in part 16. As new subcorpora are added, the documents are spread evenly over all 100 parts, such that the genres in each part remain representative of the whole corpus, making it easy to test new algorithms and tools on small but representative portions. For machine learning experiments, a proposed standard split of the GMB data is to use parts 20–99 as a training set, parts 10–19 as a development test set and parts 00–09 as a final test set. Some short-term intensive manual annotation efforts dedicated to specific annotation layers prioritise parts 00, 01, 10, 11, 20, 21 so that gold-standard data is available in each of the sets.

4.2 *The Analysis Toolchain*

The process of building the GMB takes place in a bootstrapping fashion: the raw text is first processed by a natural-language processing toolchain to produce a complete, but not necessarily fully correct first annotation. This annotation is then gradually improved using human annotation decisions. At the core of this workflow is the toolchain depicted at the bottom of Figure 4. The toolchain currently consists of four components: tokenisation, tagging, parsing, and boxing.

- **Tokenisation.** Here we use a statistical tokeniser and sentence boundary detector, Elephant, developed as part of the GMB project. The Elephant text segmentation software [27] was initially trained on a small portion of gold standard data. Manual corrections to its output (see Section 4.3) make the available amount of manually segmented text grow, and it is periodically retrained on this in order to learn new abbreviations or other tricky segmentation cases.
- **Tagging.** Various sequence taggers are employed for different annotation layers. There are currently four taggers which label individual tokens. For POS tags, we use the part-of-speech tagger included with the C&C tools, trained on CCGbank [33]. Morphological analysis is done with *morpha* [45], providing the lemma for each token. Named-entity tagging is done with the named-entity tagger included with the C&C tools, trained on the MUC data [24]. We use an in-house animacy classifier [8], which is based on a logistic regression classifier, using the implementation provided by Scikit-learn [51]. It is trained on the NXT Switchboard corpus [17] and data gathered from Wordrobe (see Section 5.2). Additionally, this classifier exploits named-entity tags, in that these override the animacy tag where applicable. That is to say, if a named entity has already been identified and tagged as, e.g., a person, this is reflected in the animacy layer with the human tag;
- **Parsing and Boxing.** The C&C syntactic parser, equipped with a model trained on CCGbank [33] produces CCG derivations. These CCG derivations are given to the semantic parser Boxer [23], providing the semantic representations used in the GMB, namely Discourse Representation Structures.

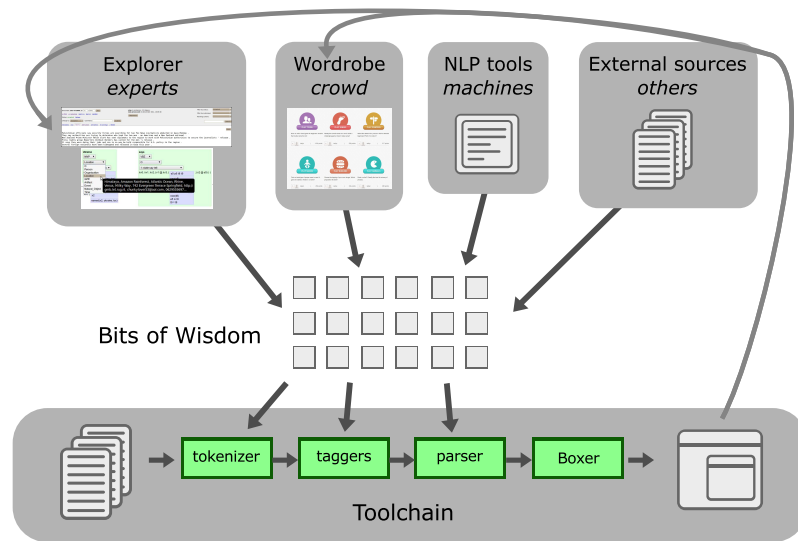


Figure 4 Graphical representation of the workflow for constructing the GMB.

4.3 *Bits of Wisdom and Automatic Adjudication*

Errors made by the NLP tools described above are unavoidable. In order to obtain reliable annotations, their output needs to be checked and corrected, ideally by human annotators. Changes in the annotation affect processing at various points within the toolchain; for example, if an annotator splits a token into two separate tokens, the part-of-speech tagger must be re-run because the new tokens cannot automatically inherit the tag of the old one. Similarly, when a part-of-speech tag is changed, the syntactic parser must be re-run because part-of-speech tags influence attachment decisions. Moreover, any change on any layer affects the final semantic representation, which thus requires re-running Boxer. It is, therefore, important to keep track of all adjustments at each step of the annotation, in order to account for their consequences at other layers.

In a traditional corpus annotation process, an annotation decision is a one-time change to a file, producing a new and better version of it. However, since the annotation process of the GMB relies on the help of a complex NLP toolchain, changing the output of a single tool once is not enough: the next time the toolchain runs, the correction would be lost. Annotation decisions therefore need to be stored and automatically applied every time the toolchain runs. Critically, adjustments to the annotation should not depend on the output of a specific tool, since this output may change due to annotations at other layers, or due to an adjustment to the tool itself. We therefore conceptualise annotation decisions as *facts* or *constraints*, rather than changes to the existing annotation. Such a constraint is called a *Bit Of Wisdom* (BOW) and contains an annotation decision that is independent from the previous an-

notation. A BOW application script checks if a machine output conforms to a BOW and if not, it makes the minimal set of changes required to make it conform. Currently, two types of BOWs are used to represent the annotation decisions made on the different layers: *segmentation* BOWs and *tag* BOWs.

The BOWs contain character offsets in order to identify the part of the raw text that the BOW provides wisdom about; they are *standoff annotations*. Each BOW is permanently stored in a relational database along with meta-information, such as its source, its creation time and the ID of the document it applies to.

Bits of wisdom are the common currency that enables wisdom from very different sources to be accumulated in the GMB in order to build the richest possible resource. The GMB uses four sources of BOWs:

1. **The wisdom of experts.** Linguistically trained annotators can use the wiki-like Explorer interface of the GMB (see Section 5.1) to make annotations. To them, the annotation process is presented as *editing* an annotated document, close to the traditional annotation process. However, behind the scenes, their edits are converted into BOWs and fed back into the toolchain when they click the “Save” button. The toolchain also allows for adding batches of BOWs, addressing systematic mistakes or inconsistencies.
2. **The wisdom of the crowd.** These BOWs are crowdsourced via a *game with a purpose* in which non-linguists collectively create BOWs (see Section 5.2).
3. **The wisdom of others.** Some sub-corpora of the GMB have already been released with linguistic annotations by others. Where the license permits, the released annotations are converted to BOWs and added to the GMB. For example, the part-of-speech annotations of the MASC corpus have been added to the corresponding GMB subcorpus in the form of BOWs.
4. **The wisdom of machines.** External NLP tools can be used even without integrating them into the toolchain, by running them on the documents and converting their output into BOWs.

Since BOWs come from different sources with varying reliability, they may conflict. The BOW application scripts therefore take the role of *judge components* that adjudicate between a set of conflicting BOWs and decide which one to apply, if any. The current strategy of this automatic adjudication process is as simple as discarding crowd BOWs if they conflict with another existing BOW, and applying the most recent remaining BOW. Future work may take confidence scores output by external tools into account.

5 Collecting Linguistic Annotations

Acquiring large amounts of reliable annotations is one of the major challenges in NLP research. As careful annotations made in a controlled environment are expensive to obtain, and cheap annotators are often unreliable, the main challenge is to

Figure 5 GMB Explorer running in a Web browser, showing navigation controls and the analysis of a document.

find a satisfactory trade-off between quantity and quality. In the Groningen Meaning Bank project we address this issue by combining different sources of annotation; annotations made in the Explorer interface are sparse but in general reliable, while crowd annotations made via a ‘Game with a Purpose’ are less informed, but numerous. This allows for a more restrictive selection procedure for the latter, in order to keep the general level of quality high. In this section we will have a closer look at both sources of input, and then discuss how they each contribute to the GMB.

5.1 Asking the Expert: Collaborative Editing

The current development version of the GMB and all changes to it are made publicly available in real time via a wiki-like Web interface, called the GMB Explorer [5].⁴ It fulfils three main functions: navigation and search through the documents, visualisation of the various levels of annotation, and manual correction of the annotations by expert annotators. Unlike most expert annotation efforts, editing is open not only to a selected team of annotators, but (after registration) to all people with the required linguistic knowledge who wish to contribute. Contributions are monitored: to date only one user submitted poor annotations; these were discarded and the user’s account was suspended. Figure 5 shows a screenshot of the Explorer interface.

The GMB Explorer interface contains basic, as well as more advanced navigational and search tools, including selection of subcorpora, word and tag searches, and a semantic lexicon (see Section 6.2). The interface shows the different levels

⁴ <http://gmb.let.rug.nl/explorer/>

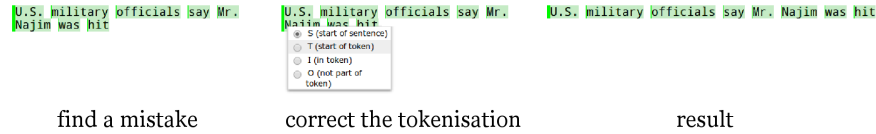


Figure 6 Three stages of correcting segmentation as shown in the GMB Explorer: (i) detecting an error, (ii) correcting the tokenisation, (iii) verifying the correction.



Figure 7 Tag edit mode with POS, lemma, animacy, senses and syntax tagging layers shown, illustrating how to adjust a POS tag.

of linguistic analysis for each document document, placed in different tabs. Basic information about the document and the raw text are shown in the *metadata* and *raw* tabs, respectively, and the *tokens* tab shows the tokenised version of the text, with one sentence per line. The *sentences* view shows the syntactic and semantic derivation for each sentence. Here, all layers of annotation can be individually shown and hidden using check-boxes. This includes CCG categories and partial, unresolved semantics on each constituent, as well as all the semantic tags for each token (named entity, thematic roles, implicit relations, animacy, scope). The *discourse* view shows a fully resolved semantic representation in the form of a DRS with rhetorical relations. Finally, there is a tab showing the *Bits of Wisdom* that have been collected for the document, and a tab containing the *warnings* produced by the NLP toolchain (if any).

The *tokenization* and *sentences* views have an “Edit” button, allowing registered users to manually correct annotations. Clicking “Edit” in the tokenisation view gives an annotator the possibility to change the IOB tags on individual characters, as Figure 6 illustrates. In the derivation view, the annotator can change tags such as part-of-speech tags and named entity tags by selecting a tag from a drop-down list (Figure 7). There is currently no way to directly edit the DRSs. This is in part by design: since the GMB adopts a lexicalised approach to constructing semantic representations, in principle, it should be possible to fix all annotation errors on the token level. In practice, this is not always true. For example, even when every token has the correct CCG category, the parser may not produce the desired representation in some cases. For such cases, or when annotators do not know how to fix a certain error, GMB Explorer provides a form for easily reporting an issue or a suggestion about a particular document to the GMB team.

| | | | | | | |
|---------------------|-----------|-------------------------|-----|-----|---------|--|
| 2013-10-31 14:20:15 | johannes | 00/0086 | BOW | tag | animacy | token 6023 (<i>work</i>) at <634,638> has animacy tag: Non-concrete |
| 2013-10-31 14:20:14 | johannes | 00/0086 | BOW | tag | animacy | token 1019 (<i>company</i>) at <127,134> has animacy tag: Organization |
| 2013-10-31 14:20:14 | johannes | 00/0086 | BOW | tag | animacy | token 2001 (<i>Halliburton</i>) at <203,214> has animacy tag: Organization |
| 2013-10-31 12:43:18 | johan.bos | 17/0053 | BOW | tag | pos | token 1002 (<i>European</i>) at <2,10> has pos tag: NNP |
| 2013-10-31 12:43:18 | johan.bos | 17/0053 | BOW | tag | ne | token 1002 (<i>European</i>) at <2,10> has ne tag: Organization |
| 2013-10-31 10:18:32 | johan.bos | 65/0693 | BOW | tok | | character at 193 labeled T: "O Hercules! |
| 2013-10-30 21:28:15 | johan.bos | 06/0691 | BOW | tag | pos | token 6012 (<i>lapsed</i>) at <591,597> has pos tag: VBD |

Figure 8 Global newsfeed of recently added BOWs in GMB Explorer.

As the updating daemon is running continually, the document is immediately re-processed after editing so that the user can directly view the new annotation with his BOW taken into account. 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. For each document, the GMB Explorer shows a time stamp indicating when it was last processed. Finally, the GMB Explorer makes the collaborative annotation process transparent through global and per-document newsfeeds of BOWs, similar to the “recent changes” and “history” feature of wikis. This is exemplified in Figure 8.

5.2 Asking the Crowd: Gamification









The idea of crowdsourcing is that some tasks that are difficult to solve for computers but easy for humans may be outsourced to a number of people across the globe. One of the prime crowdsourcing platforms is Amazon’s Mechanical Turk,⁵ an online labour marketplace where workers get paid small amounts to complete small tasks. Another crowdsourcing technique, “Game with a Purpose” (GWAP), rewards contributors with entertainment rather than money [55]. GWAPs challenge players to score high on specifically designed tasks, thereby contributing their knowledge. GWAPs were successfully pioneered in NLP by initiatives such as Phrase Detectives [20] and Play Coref [32] for anaphora resolution and ‘Jeux De Mots’ for term relations [39]. We have developed an online GWAP platform, called *Wordrobe*,⁶ which aims at collecting linguistic data for various levels of semantic annotation in the GMB.

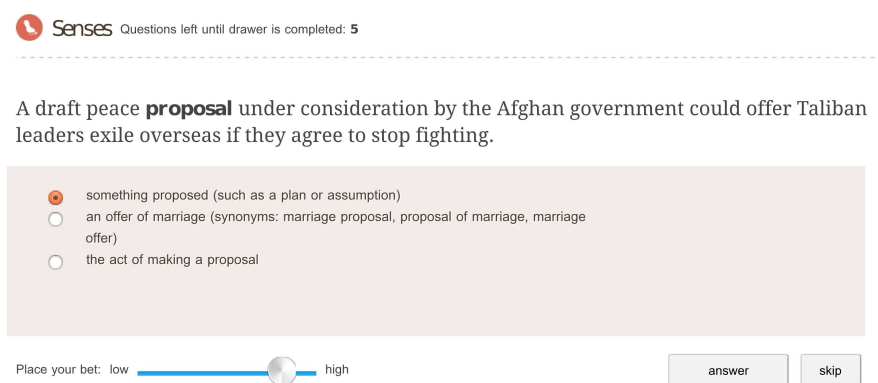
Wordrobe is a collection of games with a purpose, each targeting a specific level of linguistic annotation needed in the GMB. Current games include part-of-speech tagging, named entity tagging, co-reference resolution, word sense disambiguation, relation identification and animacy tagging. Wordrobe is designed to be used by non-experts, who can use their intuitions about language to annotate linguistic phenomena, without being discouraged by technical linguistic terminology. Therefore, the games include as few complex instructions as possible. All games share the same structure: a multiple-choice question with a small piece of text (generally one or two

⁵ <http://aws.amazon.com/mturk/>

⁶ <http://www.wordrobe.org/>

Table 6 Overview of the games provided by Wordrobe to enhance the GMB.

| Game | Task | Possible choices |
|--|-----------------------------------|--|
|  Twins | homonym disambiguation | fixed: <i>noun</i> or <i>verb</i> |
|  Senses | Word sense disambiguation | WordNet 3.1 synsets |
|  Pointers | Anaphora resolution | sequences of NNs in the context |
|  Names | Named entity tagging of NNPs | fixed class, see Sec. 3.3.1 |
|  Burgers | noun-noun compound disambiguation | prepositions |
|  Animals | Animacy classification of nouns | fixed class, see Sec. 3.3.2 |
|  Roles | Thematic role labelling | VerbNet relations |
|  Bridges | Information structure | fixed: <i>explicit</i> , <i>implicit</i> , or <i>new</i> |

**Figure 9** Screenshot from Wordrobe game *Senses*.

sentences) in which one or more words are highlighted, depending on the type of game. For each question, players can select an answer or use the skip-button to go to the next question. Players are encouraged to provide answers by means of awarded points and achievements. The points awarded are based on the agreement with other players who have provided answers to the same question, as well as a bet placed by the player, reflecting their certainty about their answer. This procedure is further detailed in [63].

All Wordrobe games consist of automatically generated multiple-choice questions from GMB documents. The choices for the questions are also automatically generated from several sources, depending on the game (see Table 6). Each question includes text extracted from the GMB with one highlighted word which the question refers to. For instance, in *Senses*, a Wordrobe game about disambiguating word senses, one word is highlighted for which the correct word sense in the given context must be selected, as shown in Figure 9. The output of the Wordrobe games are a set of BOWs.

5.3 Annotations: *Quality vs. Quantity*

Both sources of annotation presented in the previous sections have yielded a considerable amount of annotated data. Table 7 shows the number of BOWs we have collected so far, broken down by their sources. Only BOWs relative to accepted documents are considered here. Among a set of conflicting BOWs—that is, BOWs that assign different POS tags to the same token—we only count the one selected by the judge component (see Section 4.3) as “effective”, i.e. contributing to the annotation.

Table 7 Number of BOWs per type, as of November 7, 2014.

| Type | Total | Effective |
|-----------------|---------|-----------|
| expert (manual) | 44,000 | 39,279 |
| expert (script) | 134,335 | 104,744 |
| Wordrobe | 7,018 | 4,639 |
| external (MASC) | 13,351 | 9,626 |

The Wordrobe BOWs are the result of a selection procedure on player answers from the game. Since the launch of Wordrobe in September 2012, more than 1200 registered players have contributed a total of 63k single answers. In order to determine a criterion for selecting the high-quality answers among them, we conducted a first study based on the answers to the Senses game [63]. We compared several answer selection methods are based on *agreement*: in order to be reliable, the same answer should have been given to the same question by multiple different players. We compared the results of several agreement measures with gold-standard data. The results are shown in Table 8 (here, the threshold t represents the ratio between the answers for the current choice and the total number of answers to the question).

Table 8 Evaluation of selected Wordrobe choices based on different agreement measures.

| Strategy | Precision | Recall | F-score |
|---------------------------------|--------------|--------------|--------------|
| Relative majority | 0.880 | 0.834 | 0.857 |
| Absolute majority ($t = 0.5$) | 0.882 | 0.782 | 0.829 |
| Absolute majority ($t = 0.7$) | 0.945 | 0.608 | 0.740 |
| Unanimity ($t = 1$) | 0.975 | 0.347 | 0.512 |
| Chi-squared test ($p < 0.05$) | 0.923 | 0.521 | 0.666 |

The majority agreement measures are ordered according to increasing conservative-ness: the relative majority measure is least conservative, as it accepts a choice as a correct answer if most players picked the choice, whereas the measure based on unanimity only selects a choice if all six answers agree on it. The measure based on the Chi-squared test determines whether a choice is picked significantly more often

than the other choices; since the current test set only consisted of six answers per question, only choices with five or more answers were selected by this measure. The results show that given the number of answers per question in the test set, the highest F-score is obtained by using the relative majority measure. The BOWs currently obtained from Wordrobe player answers were generated using this measure. We suspect, however, that once we obtain larger amounts of answers, other more conservative measures will prove beneficial for obtaining BOWs from Wordrobe data.

6 The State of Affairs in Meaning Banking

In this final section we present the current state of the GMB project, give an overview of current research applications of the GMB, and discuss future directions of meaning banking in general and the GMB in particular.

6.1 *Availability and Distribution of the GMB*

The GMB project differs from earlier annotation work in that it regularly releases corrected and improved versions of its resources. As noted by [42], in traditional statistical NLP “there has been a very strong current against fixing data”. This has the advantage that evaluating and comparing performance of systems on original data can be done quite easily. On the other hand, even so-called gold standard annotation tends to contain mistakes and inconsistencies [42].

To get the best of both worlds, the GMB project regularly releases stable versions of its annotated corpus, via its website located at <http://gmb.let.rug.nl>. The latest stable release, version 2.2.0, comprises 10,000 texts with over a million tokens, i.e., comparable in size with the Wall Street Journal part of the PTB. The current development version contains even more texts and is accessible online through the GMB Explorer, where registered users can view the semantic annotations and contribute to the annotation process by adding BOWs.

6.2 *Applications of the GMB*

The GMB forms a rich resource of semantic information that can be used in a wide variety of language technology applications. In fact, already since its initial release in January 2012, the GMB has been adopted for research in a number of fields, including the development of algorithms for natural language generation [4], and studying quantifier scope [28], and open-domain semantic parsing [40, 7]. Learning semantic parsers from a set of language–meaning pairs, is a relatively new field and the current state-of-the-art is restricted to relatively short expressions and logical

| frequency | semantics | categories (frequency) ▼ | POS tags (frequency) | NE tag (frequency) | lemma (frequency) |
|-----------|---|--|---|--|---|
| 22302 | Av0. Av1. Av2. (v1 @ Av3. (v0 @ Av4. (e6 : (v0 @ e6)))) sLEMMA(e6) Agent(e6, v3) Theme(e6, v4) | (S[db]NP)/NP (33948) (S[ng]NP)/NP (6094) (S[pl]NP)/NP (4732) (S[dc]NP)/NP (201) (S[b]NP)/S[sem] (93) (S[pl]NP)/S[sem] (79) (S[ng]NP)/S[sem] (49) (S[ng]NP)/S[sem] (39) (S[ng]NP)/S[sem] (6) (S[ss]NP)/S[sem] (4)... | VB (33225) VBC (6099) VBN (4942) VBD (6) POS (1) NNS (1) | O (22291) (15) Time (2) (2) Person (2) Location (2) | be (973) be (458) carry (404) discuss (347) end (328) face (306) arrest (277) seek (265) reach (226)... |
| 3859 | Av0. Av1. Av2. (v1 @ Av3. (v0 @ Av4. (e6 : (v0 @ e6)))) sLEMMA(e6) Agent(e6, v3) Patient(e6, v4) | (S[b]NP)/NP (2215) (S[ng]NP)/NP (1012) (S[pl]NP)/NP (585) (S[dc]NP)/NP (47) | VB (2225) VBC (1012) VBN (616) VBD (4) NN (1) IN (1) | O (3859) | have (429) raise (204) set (186) join (192) improve (182) strengthen (108) destroy (105) expand (98) close (88) change (76)... |
| 919 | Av0. Av1. Av2. (v1 @ Av3. (v0 @ Av4. (e6 : (v0 @ e6)))) sLEMMA(e6) Theme(e6, v3) Location(e6, v4) | (S[b]NP)/NP (470) (S[ng]NP)/NP (290) (S[pl]NP)/NP (153) (S[dc]NP)/NP (5) (S[ss]NP)/NP (3) | VB (472) VBC (290) VBN (157) VBD (1) | O (919) | develop (249) form (164) open (121) issue (98) break (86) spread (40) steal (26) settle (26) stem (20) grow (17)... |

Figure 10 Excerpt from the semantic lexicon showing the most frequent entries for category $(S[dc]NP)/NP$

forms with limited expressive power [46, 68]. The GMB will form a real challenge for this area of research, with expressive meaning representations and open-domain texts.

Another side effect of producing the GMB are formal semantic lexica that can be extracted from it. As a tool for the manual study of the GMB on a higher level than individual annotations and texts, the GMB Explorer provides a web interface to the *semantic lexicon*. It shows the list of all semantic representations for individual tokens used, unique up to specific word sense symbols and specific values in numerical and time expressions. For example, Figure 10 shows the three most frequent semantic lexical entries for category $(S[dc]NP)/NP$, providing links to co-occurrence lists with particular POS tags, named-entity tags and lemmas. Note that they differ in the roles assigned to the arguments. This interface can be used to find examples of specific linguistic phenomena, to gauge their frequencies and thus to prioritise efforts for further annotation and improvement of the tools. It also gives an overview of the current inventory of lexical semantics and their dependencies on particular tagging layers.

6.3 The Future of Meaning Banking

In this chapter we introduced human-aided machine annotation, a method for developing a large semantically annotated corpus, as applied in the Groningen Meaning Bank. The method uses state-of-the-art NLP tools in combination with human input in the form of *Bits of Wisdom*. So far, we only have subjective and ad-hoc ways of measuring the quality of the semantic annotations. As the goal of the GMB is to create a gold standard for meaning representations of texts, an important direction for future work is quantifying the degree to which the gold standard is reached for a certain representation in terms of the Bits of Wisdom applied to the representation. Our working hypothesis is that the more BOWs are applied, the closer the representation reaches a gold standard.

Future work will focus on obtaining larger amounts of data, adding automated tools for detecting inconsistencies in annotation, and evaluating the annotations themselves. Moreover, this method for obtaining annotations will be applied and evaluated with respect to other linguistic phenomena, such as named entity tagging, noun-noun compound interpretation, and co-reference resolution.

We are currently preparing to add parallel texts to the corpus. This will allow us to experiment with, among other things, transferring the semantic annotation we have for English to other languages, and resolving semantic ambiguities by exploiting the fact that such ambiguities often do not overlap exactly between languages. This also requires additional annotation facilities for word and sentence alignment, and a way to align meaning representations generated by translations. Parallel meaning banking opens up a completely new series of challenges [14], such as dealing with meaning differences in translations and lexical gaps in languages. We expect that parallel meaning banking gives us new insight in formal semantic analysis, and will produce multi-lingual resources for semantic parsing. In other words: we think meaning banking is just the start of a new era in computational linguistics, and that the availability of resources like the Groningen Meaning Bank will shape research done in this field in the near future.

Acknowledgements We thank James Pustejovsky and Nancy Ide to encourage us to write this chapter. We also thank the anonymous reviewers for their valuable feedback that helped us to improve previous versions of this chapter significantly. We further would like local and visiting students who contributed to the Groningen Meaning Bank or Wordrobe: Jaap Nanninga, Jay Feldman, Lena Rampula, Hylke Postma, and Maurice Kleine. Finally we thank our crowd of expert annotators that together produced over a thousand BOWs, and the 1,580 players of Wordrobe, who all helped to improve the Groningen Meaning Bank. A final note from the authors: the ordering of the authors of this chapter is determined chronologically, reflecting the time they joined the project.

References

1. Nicholas Asher. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers, 1993.
2. Nicholas Asher and Alex Lascarides. *Logics of conversation*. Studies in natural language processing. Cambridge University Press, 2003.
3. Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Proceedings of the Conference*, pages 86–90, Université de Montréal, Montreal, Quebec, Canada, 1998.
4. Valerio Basile and Johan Bos. Aligning formal meaning representations with surface strings for wide-coverage text generation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 1–9, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.
5. Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. A platform for collaborative semantic annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 92–96, Avignon, France, 2012.

6. Valerio Basile, Johan Bos, Kilian Evang, and Noortje J. Venhuizen. Developing a large semantically annotated corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).
7. Sebastian Beschke, Yang Liu, and Wolfgang Menzel. Large-scale CCG induction from the Groningen Meaning Bank. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, 2014.
8. Johannes Bjerva. Multi-class animacy classification with semantic features. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–75, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.
9. Patrick Blackburn and Johan Bos. *Representation and Inference for Natural Language. A First Course in Computational Semantics*. CSLI, 2005.
10. Patrick Blackburn, Johan Bos, Michael Kohlhase, and Hans de Nivelle. Inference and Computational Semantics. In Harry Bunt, Reinhard Muskens, and Elias Thijsse, editors, *Computing Meaning Vol.2*, pages 11–28. Kluwer, 2001.
11. Johan Bos. Predicate Logic Unplugged. In P. Dekker and M. Stokhof, editors, *Proceedings of the Tenth Amsterdam Colloquium*, pages 133–143, ILLC/Dept. of Philosophy, University of Amsterdam, 1996.
12. Johan Bos. Implementing the Binding and Accommodation Theory for Anaphora Resolution and Presupposition Projection. *Computational Linguistics*, 29(2):179–210, 2003.
13. Johan Bos. Computational Semantics in Discourse: Underspecification, Resolution, and Inference. *Journal of Logic, Language and Information*, 13(2):139–157, 2004.
14. Johan Bos. Semantic annotation issues in parallel meaning banking. In *Proceedings of the Tenth Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-10)*, pages 17–20, Reykjavik, Iceland, 2014.
15. Johan Bos, Kilian Evang, and Malvina Nissim. Annotating semantic roles in a lexicalised grammar environment. In *Proceedings of ISA-8*, Pisa, Italy, 2012.
16. Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. Predicting the dative alternation. *Cognitive foundations of interpretation*, pages 69–94, 2007.
17. Sasha Calhoun, Jean Carletta, Jason M Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. The NXT-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419, 2010.
18. J. Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann. The NITE XML toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, 35(3):353–363, 2003.
19. Central Intelligence Agency. *The CIA World Factbook*. Potomac Books, 2006.
20. John Chamberlain, Massimo Poesio, and Udo Kruschwitz. 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, 2008.
21. Stephen Clark and James R. Curran. Parsing the WSJ using CCG and Log-Linear Models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL '04)*, pages 104–111, Barcelona, Spain, 2004.
22. Ann Copestake, Dan Flickinger, Ivan Sag, and Carl Pollard. Minimal recursion semantics: An introduction. *Journal of Research on Language and Computation*, 3(2–3):281–332, 2005.
23. James Curran, Stephen Clark, and Johan Bos. 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, 2007.
24. James R. Curran and Stephen Clark. Language Independent NER using a Maximum Entropy Tagger. In *CONLL '03 Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, pages 164–167, 2003.

25. Felice Dell’Orletta, Alessandro Lenci, Simonetta Montemagni, and Vito Pirrelli. Climbing the path to grammar: A maximum entropy model of subject/object learning. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*, pages 72–81. Association for Computational Linguistics, 2005.
26. Mike Downman, Valentin Tablan, Hamish Cunningham, and Borislav Popov. 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, 2005.
27. Kilian Evang, Valerio Basile, Grzegorz Chrupała, and Johan Bos. Elephant: Sequence labeling for word and sentence segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1426, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.
28. Kilian Evang and Johan Bos. Scope disambiguation as a tagging task. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, pages 314–320, Potsdam, Germany, March 2013. Association for Computational Linguistics.
29. Christiane Fellbaum, editor. *WordNet. An Electronic Lexical Database*. The MIT Press, 1998.
30. Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. Uby - a large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590, April 2012.
31. U. Hahn, E. Buyko, K. Tomanek, S. Piao, J. McNaught, Y. Tsuruoka, and S. Ananiadou. An annotation type system for a data-driven NLP pipeline. In *Proceedings of the Linguistic Annotation Workshop*, pages 33–40, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
32. Barbora Hladká, Jiří Mírovský, and Pavel Schlesinger. Play the language: Play coreference. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 209–212, Suntec, Singapore, August 2009. Association for Computational Linguistics.
33. J. Hockenmaier and M. Steedman. CCGbank: a corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396, 2007.
34. Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, Stroudsburg, PA, USA, 2006.
35. Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. 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, 2010.
36. Hans Kamp. A Theory of Truth and Semantic Representation. In Jeroen Groenendijk, Theo M.V. Janssen, and Martin Stokhof, editors, *Truth, Interpretation and Information*, pages 1–41. FORIS, Dordrecht – Holland/Cinnaminson – U.S.A., 1984.
37. Hans Kamp and Uwe Reyle. *From Discourse to Logic: An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT*. Kluwer, Dordrecht, 1993.
38. Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40, 2008.
39. Mathieu Lafourcade. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP’07: 7th International Symposium on Natural Language Processing*, page 7, Pattaya, Chonburi, Thailand, December 2007.
40. Phong Le and Willem Zuidema. Learning compositional semantics for open domain semantic parsing. In *Proceedings of COLING 2012*, pages 1535–1552, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.
41. Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. 2013.
42. Christopher D. Manning. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I*, pages 171–189, Berlin, Heidelberg, 2011. Springer-Verlag.

43. M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.
44. A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. The NomBank project: An interim report. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.
45. Guido Minnen, John Carroll, and Darren Pearce. Applied morphological processing of English. *Journal of Natural Language Engineering*, 7(3):207–223, 2001.
46. Raymond J. Mooney. Learning for semantic parsing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 4394 of *Lecture Notes in Computer Science*, pages 311–324. Springer Berlin Heidelberg, 2007.
47. Reinhard Muskens. Combining Montague Semantics and Discourse Representation. *Linguistics and Philosophy*, 19:143–186, 1996.
48. Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.
49. Constantin Orasan and Richard Evans. NP animacy identification for anaphora resolution. *Journal of Artificial Intelligence Research*, 29:79–103, 2007.
50. Martha Palmer, Paul Kingsbury, and Daniel Gildea. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.
51. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
52. Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, 2002.
53. Christopher Potts. *The Logic of Conventional Implicatures*. Oxford University Press, USA, 2005.
54. Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. The Penn Discourse TreeBank as a resource for natural language generation. In *Proc. of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32, 2005.
55. James Pustejovsky and Amber Stubbs. *Natural Language Annotation and Machine Learning*. O’Reilly Media, 2012.
56. Anette Rosenbach. Animacy and grammatical variation—findings from English genitive variation. *Lingua*, 118(2):151–171, 2008.
57. Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *LREC*, 2002.
58. Mark Steedman. *The Syntactic Process*. The MIT Press, 2001.
59. Anatol Stefanowitsch. Constructional semantics as a limit to grammatical alternation: The two genitives of English. *TOPICS IN ENGLISH LINGUISTICS*, 43:413–444, 2003.
60. Jörg Tiedemann. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, 2009.
61. Rob A. Van der Sandt. Presupposition Projection as Anaphora Resolution. *Journal of Semantics*, 9:333–377, 1992.
62. Jan van Eijck and Hans Kamp. Representing Discourse in Context. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pages 179–240. Elsevier, MIT, 1997.
63. Noortje J. Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. Gamification for word sense labeling. *Proc. 10th International Conference on Computational Semantics (IWCS-2013)*, pages 397–403, 2013.

64. Noortje J. Venhuizen, Johan Bos, and Harm Brouwer. Parsimonious semantic representations with projection pointers. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 252–263, Potsdam, Germany, 2013. Association for Computational Linguistics.
65. Noortje J. Venhuizen, Johan Bos, Petra Hendriks, and Harm Brouwer. How and why conventional implicatures project. In *Proceedings of the 24th Semantics and Linguistic Theory Conference (SALT 24)*, pages 63–83, New York, May 30 – June 1 2014. New York University.
66. Adam Wyner, Johan Bos, Valerio Basile, and Paulo Quaresma. An empirical approach to the semantic representation of laws. In *JURIX*, pages 177–180, 2012.
67. Annie Zaenen, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M Catherine O’Connor, and Tom Wasow. Animacy encoding in english: why and how. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 118–125. Association for Computational Linguistics, 2004.
68. Luke Zettlemoyer and Michael Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 658–666, Arlington, Virginia, 2005. AUAI Press.