# Thirty Musts for Meaning Banking

**Johan Bos**
Center for Language and Cognition
University of Groningen
johan.bos@rug.nl

**Lasha Abzianidze**
Center for Language and Cognition
University of Groningen
l.abzianidze@rug.nl

## Abstract

Meaning banking—creating a semantically annotated corpus for the purpose of semantic parsing or generation—is a challenging task. It is quite simple to come up with a complex meaning representation, but it is hard to design a simple meaning representation that captures many nuances of meaning. This paper lists some lessons learned in nearly ten years of meaning annotation during the development of the Groningen Meaning Bank (Bos et al., 2017) and the Parallel Meaning Bank (Abzianidze et al., 2017). The paper's format is rather unconventional: there is no explicit related work, no methodology section, no results, and no discussion (and the current snippet is not an abstract but actually an introductory preface). Instead, its structure is inspired by work of Traum (2000) and Bender (2013). The list starts with a brief overview of the existing meaning banks (Section 1) and the rest of the items are roughly divided into three groups: corpus collection (Section 2 and 3, annotation methods (Section 4–11), and design of meaning representations (Section 12–30). We hope this overview will give inspiration and guidance in creating improved meaning banks in the future.

## 1 Look at other meaning banks

Other semantic annotation projects can be inspiring, help you to find solutions to hard annotation problems, or to find out where improvements to the state of the art are still needed (Abend and Rappoport, 2017). Good starting points are the English Resource Grammar (Flickinger, 2000, 2011), the Groningen Meaning Bank (GMB, Bos et al. 2017), the AMR Bank (Banarescu et al., 2013), the Parallel Meaning Bank (PMB, Abzianidze et al. 2017), Scope Control Theory (Butler and Yoshimoto, 2012), UCCA (Abend and Rappoport, 2013), Prague Semantic Dependencies (Hajič et al., 2017) and the ULF Corpus based on Episodic Logic (Kim and Schubert, 2019). The largest differences between these approaches can be found in the expressive power of the meaning representations used. The simplest representations correspond to graphs (Banarescu et al., 2013; Abend and Rappoport, 2013); slightly more expressive ones correspond to first-order logic (Oepen et al., 2016; Bos et al., 2017; Abzianidze et al., 2017; Butler and Yoshimoto, 2012), whereas others go beyond this (Kim and Schubert, 2019). Generally, an increase of expressive power causes a decrease of efficient reasoning (Blackburn and Bos, 2005). Semantic formalisms based on graphs are attractive because of their simplicity, but will face issues when dealing with negation in inference tasks (Section 21). The choice might depend on the application (e.g., if you are not interested in detecting contradictions, coping with negation is less important), but arguably, an open-domain meaning bank ought to be independent of a specific application.

## 2 Select public domain corpora

Any text could be protected by copyright law and it is not always easy to find suitable corpora that are free from copyright issues. Indeed, the relationship between copyright of texts and their use in natural language processing is complex (Eckart de Castilho et al., 2018). Nonetheless, it pays off to make some effort by searching for corpora that are free or in the public domain (Ide et al., 2010). This makes it easier for other researchers to work with it, in particular those that are employed by institutes with lesser financial means. The GMB only includes corpora from the public domain (Basile et al., 2012b). Free parallel corpora are also available via OPUS (Skadiņš et al., 2014). Other researchers take advantage of vague legislation and



*Proceedings of the First International Workshop on Designing Meaning Representations*, pages 15–27

distribute corpora quoting the right of fair use (Postma et al., 2018). Recently, crowd sourcing platforms such as Figure Eight make datasets available, too ("Data For Everyone"), under appropriate licensing. While targeting the public domain corpora, one might need to bear in mind the coverage of the corpora depending on the objectives of semantic annotation.

## 3 Freeze the corpus before you start

Once you start your annotation efforts, it is a good idea to freeze the corpora that will comprise your meaning bank.[1] In the GMB project (Basile et al., 2012b), the developers were less strict in maintaining this principle. During the project they came across new corpora, but after adding them to the GMB they were forced to fix and validate annotations on many levels to get the newly added corpus up to date and in sync with the rest. This problems manifests itself especially for corpora that are constructed via a phenomenon-driven annotation approach (Section 24).

## 4 Work with raw texts in your corpus

Keep the original texts as foundation for annotation. Never ever carry out any semantic annotation on tokenised texts, but use stand-off annotation on character offsets (Section 5). Tokenisation can be done in many different ways, and the *atoms of meaning* certainly do not correspond directly to words. Most of the current conventions in tokenisation are based on what has been used in (English) syntax-oriented computational linguistics and can be misleading when other languages are taken into consideration (Section 29). Moreover, if you use an off-the-shelf tokeniser, you will find out soon that it makes mistakes—and correcting those would break any annotations done at the word token level. More likely, during your annotation project, you will find the need to change the tokenisation guidelines to deal properly with multi-word expressions (Section 22). In addition, punctuation and spacing carry information that could be useful for deep learning approaches, and their original appearance should therefore in one way or another should be preserved. An example: a "New York-based" company could be a new company based in York, but the other interpretation is more likely. In an NLP-processing pipeline, it is too late for syntax to fix this in a compositional way—the tokenisation needs to be improved.

## 5 Use stand-off annotation

Stand-off annotation is a no-brainer as it offers a lot more flexibility. It enables keeping annotations separate from the original raw text, where ideally each annotation layer has its own file (Ide and Romary, 2006; Pustejovsky and Stubbs, 2012). It is best executed with respect to the character offsets of the raw texts in the corpus (Section 4). A JSON or XML-based annotation file can always be generated from this, should the demand be there. Stand-off annotation is in particular advantageous in a setting where several layers of annotation interact with each other (typically in a pipeline architecture). This was extremely helpful in the GMB (Bos et al., 2017) where the document segmentation (sentence and word boundaries) got improved several times during the project, without having any negative effect on annotation occurring later in the semantic processing pipeline (such as part-of-speech tagging and named entity recognition).

## 6 Consider manual annotation

Several meaning banks are created with the help of a grammar. The best example here is the sophisticated English Resource Grammar (Flickinger, 2000, 2011) used to produce the treebanks, Redwoods (Oepen et al., 2004) and DeepBank (Flickinger et al., 2012), annotated with English Resource Semantics (ERS) in a compositional way, by letting the annotator pick the correct or most plausible analysis. Similarly, the meaning representations in the GMB are system-produced and partially hand-corrected (Bos et al., 2017), using a CCG parser (Clark and Curran, 2004). Likewise, the meaning representations in the PMB are system-produced with the help of a CCG parser (Lewis and Steedman, 2014) and some of it is completely hand-corrected. In contrast, the meaning representations of the AMR Bank are completely manually manufactured—without the aid of a grammar—with the help of an annotation interface and an extensive manual (Banarescu et al., 2013). Bender et al. (2015) argue that grammar-based meaning banking requires less annotation guidelines, that it provides more consistent anal-

[1] Freezing the corpora already fixes certain data statements for your meaning bank, like curation rationale, language variety, and text characteristics. Communicating these data statements is important from an application point of view (Bender and Friedman, 2018).

yses, and that it is more scalable. The downside of grammar-based annotation is that several compound expressions are not always compositional (negative and modal concord, postnominal genitives ("of John's"), odd punctuation conventions, idioms), and that grammars with high recall and precision are costly to produce (the impressive English Resource Grammar took about several years to develop, but it is restricted to just one language).

## 7 Make a friendly annotation interface

Annotation can be fun (especially if gamification is applied, see Section 9), but it can also be tedious. A good interface helps the annotator to make high-quality annotations, to work efficiently, and to be able to focus on particular linguistic phenomena. An annotation interface should be web-based (i.e., any browser should support it), simple to use, and personalised.[2] The latter grants control over annotations of particular users. The "Explorer" (Basile et al., 2012a) introduced in the GMB and later further developed in the PMB, has various search abilities (searches for phrases, regular expressions, and annotation labels), a statistics page, a newsfeed, and a user-friendly way to classify annotations as "gold standard". The inclusion of a "sanity checker" helps to identify annotation mistakes, in particular if there are several annotation layers with dependencies. It is also a good idea to hook the annotation interface up with a professional issue reporting system.

## 8 Include an issue reporting system

Annotators will sooner or later raise issues, have questions about the annotation scheme, or find bugs in the processing pipeline. This is valuable information for the annotation project and should not get lost. The proper way to deal with this is to include a sophisticated bug reporting system in the annotation interface. For the GMB (Bos et al., 2017) and the PMB (Abzianidze et al., 2017), the Mantis Bug Tracker[3] was incorporated inside the Explorer (Basile et al., 2012a). Besides Mantis there are many other free and open source web-based bug tracking systems available. A bug tracker enables one to categorize issues, assign them to team members, have dedicated discussion thread for each issue, and keep track of all improvements made in a certain time span (useful for the documentation in data releases).

## 9 Be careful with the crowd

Following the idea of Phrase Detectives (Chamberlain et al., 2008), in the GMB (Bos et al., 2017) a game with a purpose (GWAP) was introduced to annotate parts of speech, antecedents of pronouns, noun compound relations (Bos and Nissim, 2015), and word senses (Venhuizen et al., 2013). The quality of annotations harvested from gamification was generally high, but the amount of annotations relatively low—it would literally take years to annotate the entire GMB corpus. An additional problem with GWAPs is recruiting new players: most players play the game only once, and attempts to make the game addictive could be irresponsible (Andrade F.R.H., 2016). The alternative, engaging people by financially awarding them via crowdsourcing platforms such as Mechanical Turk or Figure Eight, solves the quantity problem (Pustejovsky and Stubbs, 2012), but introduces other issues including the question what a proper wage would be (Fort et al., 2011) and dealing with tricksters and cheaters (Buchholz and Latorre, 2011).

## 10 Profit from lexicalised grammars

A lexicalised grammar gives an advantage in annotating syntactic structure. In case of the compositional semantics, this also leads to automatic construction of the phrasal semantics. This is because, in a lexicalised grammar, most of the grammar work is done in the lexicon (there is only a dozen general grammar rules), and annotation is just a matter of giving the right information to a word (rather than selecting the correct interpretation from a possibly large set of parse trees). In the PMB a lexicalised grammar is used: Combinatory Categorial Grammar (CCG, Steedman 2001), and the core annotation layers for each word token are a CCG category, a semantic tag (Abzianidze and Bos, 2017), a lemma, and a word sense. Annotating thematic roles (Section 18) is also convenient in a lexicalised grammar environment (Bos et al., 2012). Finally, a lexicalised grammar coupled with compositional semantics facilitates annotation projection for meaning preserving translations and opens the door to multilingual meaning banking (Section 29). Projection of meaning representation from one sentence to another is reduced to word alignment and word-level annota-

[2] For more details about web-based collaborative annotation tools we refer to Biemann et al. (2017).

[3] https://www.mantisbt.org/

tion transfer. This type of projection is underlying the idea of moving from the monolingual GMB to the multilingual PMB.

## 11 Try to use language-neutral tools

Whenever possible, in machine-assisted annotation, get language technology components that are not tailored to specific languages, because this increases portability of meaning processing components to other languages (Section 29). The statistical tokeniser (for word and sentence segmentation) used in the PMB is Elephant (Evang et al., 2013). The current efforts in multi-lingual POS-tagging, semantic tagging (Abzianidze and Bos, 2017) and dependency parsing are promising (White et al., 2016). In the PMB a categorial grammar is used to cover four languages (English, Dutch, German, and Italian), using the same parser and grammar, but with language-specific statistical models trained for the EasyCCG parser (Lewis and Steedman, 2014). Related are grammatical frameworks designed for parallel grammar writing (Ranta, 2011; Bender et al., 2010).

## 12 Apply normalisation to symbols

Normalising the format of non-logical symbols (the predicates and individual constants, as opposed to logical symbols such as negation and conjunction) in meaning representations decreases the need for awkward background knowledge rules that would otherwise be needed to predict correct entailments. Normalisation (van der Goot, 2019) can be applied to date expressions (e.g., the 24th of February 2010 vs. 24-02-2010 or dozens of variations on these), time expressions (2pm, 14:00, two o'clock), and numerical expressions (twenty-four, 24, vierundzwanzig; three thousand, 3,000, 3000, 3 000). Compositional attempts to any of the above mentioned classes of expressions are highly ambitious and not recommended. Take, for instance, the Dutch clock time expression "twee voor half vier", which denotes 03:28 (or 15:28)—how would you derive this compositionally in a computational straightforward way? Other normalisations for consideration are expansion of abbreviations to their full forms, lowercasing proper names, units of measurement, and scores of sports games. To promote inter-operability between annotated corpora, it is a good idea to check whether any standards are proposed for normalisation (Pustejovsky and Stubbs, 2012).

## 13 Limit underspecification

Underspecification is a technique with the aim to free the semantic interpretation component from a disambiguation burden (Reyle, 1993; Bos, 1996; Copestake et al., 2005). In syntactic treebanks, however, the driving force has been to assign the most plausible parse tree to a sentence. This makes sense for the task of statistical (syntactic) parsing. The same applies to (statistical) semantic parsing: a corpus with the most likely interpretation for sentences is required. Moreover, it is not straightforward to draw correct inferences with underspecified meaning representations (Reyle, 1995). So it makes sense, at least from the perspective of semantic annotation, to produce the most plausible interpretation for a given sentence. Consider the following examples. A "sleeping bag" could be a bag that is asleep, but it is very unlikely (even in a Harry Potter setting), so should be annotated as a bag designed to be slept in. In the sentence "Tom kissed his mother", the possessive pronoun could refer to a third party, but by far the most likely interpretation is that Tom's mother is kissed by Tom, and that reading should be reflected in the annotation. Genuine scope ambiguities are relatively rare in ordinary text, and it is questionable whether the representational overhead of underspecified scope is worth the effort given the low frequency of the phenomenon. Nonetheless, resolving ambiguities is sometimes hard, in particular for sentences in isolation. What is plausible for one annotator is implausible for another. Finally, one needs to be careful, as annotation guidelines that give preference for one particular reading (based on statistical plausbility) have the danger of introducing or even amplifying bias.

## 14 Beware of annotation bias

Assigning the most likely interpretation to a sentence can also give an unfair balance to stereotypes. In the PMB, gender of personal proper names are annotated. In many cases this is a straightforward exercise. But there are sometimes cases where the gender of a person is not known. The disturbing distribution of male versus female pronouns (or titles) strongly suggests that a female is the least likely choice (Webster et al., 2018). But following this statistical suggestion only causes

greater divide. The PMB annotation guidelines for choosing word senses (Secion 15) are such that when it is unclear what sense to pick, the higher sense (thus, the most frequent one), must be selected. This is bad, because systems for word sense disambiguation already show a tendency towards assigning the most frequent sense (Postma et al., 2016). More efforts are needed to reduce bias (Zhao et al., 2017).

## 15 Use existing resources for word senses

The predicate symbols that one finds in meaning representation are usually based on word lemmas. But words have no interpretation, and a link to concepts in an existing ontology (Lenat, 1995; Navigli and Ponzetto, 2012) is something that is needed to make the non-logical symbols in meaning representations interpretable. In the AMR Bank, verbs are disambiguated by OntoNotes senses (Banarescu et al., 2013). In the PMB, nouns, verbs, adjectives and adverbs are labelled with the senses of (English) WordNet (Fellbaum, 1998). Picking the right sense is sometimes hard for annotators, sometimes because there is too little context, but also because the definitions of fine-grained senses are sometimes hard to distinguish from each other (Lopez de Lacalle and Agirre, 2015). Annotation guidelines are needed for ambiguous cases where syntax doesn't help to disambiguate: "Swimming is great fun." (`swimming.n.01` or perhaps `swim.v.01`?), "Her words were emphasized." (`emphasized.a.01` or `emphasize.v.02`?). WordNet's coverage is impressive and substantial, but obviously not all words are listed (example: names of products used as nouns) and sometimes it is inconsistent (for instance, "apple juice" is in WordNet, but "cherry juice" is not). Many WordNets exists for languages other than English (Navigli and Ponzetto, 2012; Bond and Foster, 2013).

## 16 Apply symbol grounding

Symbol grounding helps to connect abstract representations of meaning with objects in the real world or to unambiguous descriptions of concepts or entities. This happens on the conceptual level with mapping words to WordNet synsets or to a well-defined inventory of relations. Princeton WordNet (Fellbaum, 1998) lists several instances of famous persons but obviously the list is incomplete. The AMR Bank includes links from named entities to wikipedia pages, but obviously not every named entity has a wikipedia entry. To our knowledge, no other meaning banks apply wikification. Other interesting applications for symbol grounding are GPS coordinates for toponyms (Leidner, 2008), visualisation of concepts or actions (Navigli and Ponzetto, 2012), or creating timelines (Bamman and Smith, 2014).

## 17 Adopt neo-Davidsonian events

It seems that in most (if not all) semantically annotated corpora a neo-Davidsonian event semantics is adopted. This means that every event introduces its own entity as a variable, and this variable can be used to connect the event to its thematic roles. In the original Davidsonian approach, an event variable was simply added to the predicate introduced by the verb (Davidson, 1967; Kamp and Reyle, 1993) as a way to add modifiers (e.g., moving from `eat(x,y)` to `eat(e,x,y)` for a transitive use of *to eat*). In most modern meaning representations thematic roles are introduced to reduce the number of arguments of verbal predicates to one, also known as the neo-Davidsonian tradition (Parsons, 1990) (e.g., moving from `eat(e,x,y)` to `eat(e) AGENT(e,x) PATIENT(e,y)`). A direct consequence of a neo-Davidsonion design is the need for an inventory of thematic roles. But there is also an alternative, which is given a fixed arity to event predicates, of which some of them may be unused (Hobbs, 1991) when the context does not provide this information (e.g., for the intransive usage of *to eat*, still maintain `eat(e,x,y)` where `y` is left unspecified).

## 18 Use existing role labelling inventories

A neo-Davidsonian approach presupposes a dictionary of thematic (or semantic) role names. There are three popular sets available: PropBank, VerbNet, and FrameNet. PropBank (Palmer et al., 2005) proposes a set of just six summarising roles: ARG0 (Agent), ARG1 (Patient), ARG2 (Instrument, Benefactive, Attribute), ARG3 (Starting Point), ARG4 (Ending Point), ARGM (Modifier). The interpretation of these roles are in many cases specific to the event in which they participate. The AMR Bank adopts these PropBank roles (Banarescu et al., 2013). VerbNet has a set of about 25 thematic roles independently defined from the verb classes (Kipper et al., 2008). A few examples are: Agent, Patient, Theme, Instru-

ment, Experiencer, Stimulus, Attribute, Value, Location, Destination, Source, Result, and Material. The PMB adopts the thematic roles of VerbNet. FrameNet is organised quite differently. Its starting point is not rooted in linguistics, but rather in real-world situations, classified as frames (Baker et al., 1998). Frames have frame elements that can be realised by linguistic expressions, and they correspond to the PropBank and VerbNet roles. There are more than a thousand different frames, and each frame has its own specific role set (frame elements). For instance, the Buy-Commerce frame has roles Buyer, Goods, Seller, Money, and so on. There are also recent proposals for comprehensive inventories for roles introduced by prepositional and possessive constructions (Schneider et al., 2018). In the PMB, we employ a unified inventory of thematic roles (an extension of the VerbNet roles) that is applicable to verbs, adjectives, prepositions, possessives or noun modifiers.

## 19 Treat agent nouns differently

Agent and recipient nouns (nouns that denote persons performing or receiving some action, such as employee, victim, teacher, mother, cyclist, victim) are intrinsically relational (Booij, 1986). Modelling them like ordinary nouns, i.e., as one-place predicates, can give rise to contradictions for any individual that has been assigned more than one role, because while you want to be able to state that a violin player is not the same thing as a mother, a person could perfectly be a mother and a violin player at the same time. Moreover, a fast cyclist could be a slow driver. Incorrect modeling can furthermore lead to over-generation of some unmanifested relations (for instance, if Butch is Vincent's boss and Mia's husband, a too simple model would predict that Butch is also Vincent's husband and Mia's boss. In the AMR Bank (Banarescu et al., 2013) agent nouns are decomposed (e.g., an "investor" is a person that invests). In the PMB agent nouns introduce a mirror entity (e.g. an "investor" is a person with the role of investor).

## 20 Beware of geopolitical entities

Names used to refer to geopolitical entities (GPEs) are a real pain in the neck for semantic annotators. How many times did we change the annotation guidelines for these annoying names! The problem is that expressions like "New York", "Italy", or "Africa" can refer to locations, their governments, sport squads that represent them, or the people that live there (and in some case to multiple aspects at the same time, as in "Italy produces better wine than France"). This instance of systematic polysemy manifests itself for all classes of GPE, including continents, countries, states, provinces, cities, and so on. Detailed instructions for annotating GPEs can be found in the ACE annotation guidelines (Doddington et al., 2004).

## 21 Give scope to negation

Sentence meaning is about assigning truth conditions to propositions (Section 23). Negation plays a crucial role here—in fact, the core of semantics is about negation, identifying whether a statement is true of false. Negation is a semantic phenomenon that requires scope, in other words, it cannot be modelled by simply applying it as a property of an entity. It should be clear—explicit or implicit—what the scope of any negation operator is, i.e. the parts of the meaning representation that are negated. The GMB, PMB and DeepBank (Flickinger et al., 2012) assign proper scope to negation (the latter with the help of underspecification). In AMR Bank negation is modelled with the help of a relation, and this doesn't always get the required interpretation (Bos, 2016). Negation can be tricky: negation affixes (Section 23) require special care, negative concord (Section 6) and neg raising (Liu et al., 2018) are challenges for compositional approaches to meaning construction.

## 22 Pay attention to compound words

In the GMB (Bos et al., 2017) we largely ignored multi-word expressions (MWEs), believing that compositionality would eventually do away with it. Except it doesn't. MWEs come in various forms, and require various treatments (Sag et al., 2002). Think about proper names (names of persons, companies, locations, events), titles and labels (of people, of books, chapters, of songs), compounds, phrasal verbs, particle verbs, fixed phrases, and idioms. Consider for instance "North and South Dakota", it is quite a challenge to derive the representation state(x) & name(x,'North-Dakota') in a compositional way. And many compounds are not compositional ("peanut butter" is not butter, and "athlete's foot" is not a body part but a nasty infection). It is hard to decide where to draw the line between a compositional and non-compositional approach to multi-

word expressions. Even though "red wine" is written in English with two words, in German it is written in one word ("rotwein"). WordNet (Fellbaum, 1998) lists many multi-word expressions and could be used as a resource to decide whether a compounds is analysed compositionally or not. In the PMB, titles of songs or other artistic works are treated as a single token (because they are proper names), which works fine for "Jingle Bells" but becomes a bit awkward and uncomfortable with longer titles such as Lennon and McCartney's "Lucy in the Sky with Diamonds", or Pink Floyds's "Several Species of Small Furry Animals Gathered Together in a Cave and Grooving With A Pict". It is quite unfair and unrealistic to expect the tokeniser to recognise this as a multi-word expression. The alternative, applying some reinterpretation after having first carried out a compositional analysis, puts a heavier burden on the syntax-semantics interface. The bottom line is that MWEs form a wild bunch of expressions for which a general modelling strategy covering all types does not seem to exist. There also seems to be a connection with quotation (Maier, 2014).

## 23 Use inference tests in design

The driving force to motivate how to shape or what to include in a meaning representation should be textual entailment or contradiction checks (this is a practice borrowed from formal semantics). For instance, when designing a meaning representation for adjectives, the meaning for "ten-year-old boy" should not imply that the boy in question is old. Likewise, the meaning representation for "unhappy" should not be the same as that for "not happy", because the meanings of these expressions are not equivalent (as "Bob is not happy" doesn't entail "Bob's unhappy"—Bob can be both not happy and not unhappy—even though the entailment holds in the reverse direction: if Bob is unhappy, he is not happy). Similarly, the meaning representation for "Bologna is the cultural capital of Italy" should not lead to the incorrect inference that "Bologna is the capital of Italy". In addition, or as alternative to inference checks, is applying the method of model-theoretic interpretation (Blackburn and Bos, 2005) when designing meaning representations. It should be clear what a representation actually means, in other words, under which conditions it is true or false. A formal way of defining this is via models of situation, and a satisfaction definition that tells us, given a certain situation, whether a statement holds or doesn't. This method was introduced by the logician Tarski (Tarski and Vaught, 1956). It bears similarities with posing a query to a relational database. The method forces you to make a strict distinction between logical (negation, disjunction, equality) and non-logical symbols (the predicates and individual constants in your meaning representation).

## 24 Divide and conquer

Do not try to do model all semantic phenomena the first time around. There are just too many. Some good candidates to put on hold are plurals, tense, aspect, focus, presupposition (see Section 25), and generics (more in Section 27), because a proper treatment of these phenomena requires a lot more than a basic predicate-argument structure. A strict formalisation of plurals quickly leads to complicated representations (Kamp and Reyle, 1993), leading to compromising approximations in the AMR Bank (Banarescu et al., 2013) or PMB (Abzianidze et al., 2017). In the GMB (Bos et al., 2017) and the AMR Bank tense is simply ignored. Annotating aspect is complex—for instance, the use of the perfect differs enormously even between closely related languages such as English, Dutch, and Italian (van der Klis et al., 2017). These complications lead to a simple annotation model in the PMB where tense is reduced to a manageable set of three tenses: past, present and future. There are, therefore, a lot of interesting problems left for the second round of semantic annotation!

## 25 Put complex presuppositions on hold

Presuppositions are propositions that are taken for granted. Several natural language expressions introduce presuppositions. These expressions are called presuppositions triggers. (For instance, "Mary left, too." presupposes that someone else besides Mary left. Here "too" is the trigger of this presupposition.) There are many different kinds of triggers, and many do not contribute to the meaning of the sentence, but rather put constraints on the context. The question, then, is what to do with them in a meaning banking project. Some classes of presupposition triggers, referring expressions including proper names, possessive phrases, and definite descriptions, can be treated in a similar way as pronouns, as is done in the GMB and

the PMB, following Bos (2003). Yet there are other classes of triggers that are notoriously hard to represent, because they require some "copying" of large pieces of meaning representation, interact with focus, and require non-trivial semantic composition methods. To these belong implicative verbs (manage), focusing adverbs (only, just), and repetition particles (again, still, yet, another). For instance, although in the PMB a sentence like "The crowd applauded again." is the presupposition trigger, "again" is semantically tagged as a repetition trigger, for now it doesn't perform any costly operations on the actual meaning representation. The first alternative, a meaning representation with two different applauding events that are temporally related, is complicated to construct. The second alternative, introducing "again" as a predicate, doesn't make sense semantically (what is the meaning of "again"?), or as an operator (again, how will it be defined logically?) isn't attractive either. There are, currently no good ways to deal with complex presupposition triggers, and more research is needed here turning formal ideas (Kamp and Rossdeutscher, 2009) into practical solutions.

## 26 Respect elliptical expressions

They are invisible, but omnipresent: elliptical expressions. Comparative ellipsis is present in many languages ("My hair is longer than Tom's"). In English, verb phrase ellipsis occurs ("Tom eats asparagus, but his brother doesn't."), which is well studied (Dalrymple et al., 1991), and annotated corpora exist as well (Bos and Spenader, 2011). Dutch and German exhibit a large variety of gapping cases ("Tom isst Spargel, aber sein Bruder nicht."). Italian is a language with pro-drop ("Ho fame", i.e., (I) am hungry). Ellipsis requires a dedicated component in a pipeline architecture. In the PMB the inclusion of an ellipsis layer has been postponed for the benefit of other components, features, and efforts. As a consequence, a growing number of documents cannot be added to the gold set because there isn't an adequate way of dealing with a missing pronoun, an odd comparison expression, or an elided verb phrase.

## 27 Think about generics

Generic statements and habituals are hard to model straightforwardly in first-order logic (Carlson, 1977). The sentence "a lion is strong" or "a dog has four legs" is not about a particular lion or dog, nor is it about all dogs or lions. The inventor of "the typewriter" was not the inventor of a particular typewriter, but of the typewriter concept in general. Such generic concepts are also known as *kinds* in the literature (Reiter and Frank, 2010). It is not impossible to approximate this in first-order logic, but it requires an ontological distinction between entities denoting individuals and entities denoting concepts (kinds). A further question is how tense should be annotated in habitual sentences, as in "Jane used to swim every day" (in some period in the past, Jane swam every day) or "Jane swims every day" (in the current period, Jane swims every day). To our knowledge, none of the existing meaning banks have a satisfactory treatment of generics, even though techniques have been proposed to detect generics (Reiter and Frank, 2010; Friedrich and Pinkal, 2015). Recent proposals try to change this situation (Donatelli et al., 2018).

## 28 Don't try to be clever

The English verb "to be" (and its counterpart in other Germanic languages) is a semantic nuisance. When used as an auxiliary verb—including predicative usages of adjectives—there isn't much to worry about it, as it only semantically contributes tense information. However, when used as a copula it can express identity, locative information, or predications involving nouns. From a logical perspective, it might seem attractive to use equality in these cases and interpret "to be" logically rather than lexically, (Blackburn and Bos, 2005), but this makes it impossible to include tense information, unless equality is (non-standardly) viewed as a three-place relation. There are various senses for "be" in WordNet, and it makes pragmatically sense to use these: "This is a good idea" (sense 1), "John is the teacher" (sense 2), "the book is on the table" (sense 3), and so on. A similar story can be told for "to have" in expressions like "Mary has a son", where the first attempt in the PMB was to analyse "to have" in such possessive constructions as logical, i.e. only introducing tense information, and coerce the relational noun "son" into a possessive interpretation. This was soon abandoned due to complications in composition.

## 29 Don't focus on just one language

Most meaning banks consider just one language, and usually this is English. This is understand-

able, as English is the current scientific language, but it is also risky, because when designing meaning representation decisions could be made that work for English but not for other languages. Phenomena such as definite descriptions, ellipsis, possessives, aspect, and gender, behave even in closely related languages quite differently from each other. Dealing with multiple languages is, without any doubt, harder, but if one takes several languages into account at the same time the result is more likely to be more language-neutral meaning representations. And that's what meanings should be, they are abstract objects, independent of the language used to expressed them. Of course, there are concepts that can be expressed in certain languages with a single word that other languages are not capable of, but the core of meaning representations should be agnostic to the source language. A good starting point is to work with typologically-related languages. An efficient annotation technique to cover multiple languages is *annotation projection* (Evang and Bos, 2016; Liu et al., 2018). This requires a parallel corpus and automatic word alignment, and existing semantic annotations for at least one language.

## 30 Measure meaning discrepancies

A large part of the users of semantically annotated corpora are from the semantic parsing area, and they need to be able to measure and quantify their output with respect to gold standard meanings. The currently accepted methods are based on precision and recall on the components of the meaning representation by converting them to triples or clauses (Allen et al., 2008; Dridan and Oepen, 2011; Cai and Knight, 2013; Van Noord et al., 2018; Kim and Schubert, 2019). In a parallel corpus setting, such evaluation measures can also be used to compare the meaning representation of a source text and its translation (Saphra and Lopez, 2015). This is done in the PMB, where a non-perfect meaning match between source and target helps the annotator to identify possible culprits. It is important to note that most of these matching techniques check for syntactic equivalence, and don't take semantic equivalence into account—the same meaning could be expressed by syntactically different representations. The approach by Van Noord et al. (2018) applies normalisation steps for word senses to make matching more semantic.

## Acknowledgments

We would like to thank the two anonymous reviewers for their comments—they helped to improve this paper considerably. Reviewer 1 gave us valuable pointers to the literature that we missed, and spotted many unclear and ambiguous formulations. Reviewer 2 was disappointed by the first version of this paper— we hope s/he likes this improved version better. This work was funded by the NWO-VICI grant Lost in Translation Found in Meaning (288-89-003).

## References

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (ucca). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Omri Abend and Ari Rappoport. 2017. The state of the art in semantic representation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 77–89, Vancouver, Canada. Association for Computational Linguistics.

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 242–247, Valencia, Spain.

Lasha Abzianidze and Johan Bos. 2017. Towards universal semantic tagging. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017) – Short Papers*, pages 1–6, Montpellier, France.

James F. Allen, Mary Swift, and Will de Beaumont. 2008. Deep Semantic Analysis of Text. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 343–354. College Publications.

Isotani S. Andrade F.R.H., Mizoguchi R. 2016. The bright and dark sides of gamification. *Intelligent Tutoring Systems. ITS 2016. Lecture Notes in Computer Science*, 9684.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. 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.

David Bamman and Noah A. Smith. 2014. Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics*, 2:363–376.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012a. 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.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Joost Venhuizen. 2012b. Developing a large semantically annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Emily M. Bender. 2013. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation*, 8(1):23–72.

Emily M. Bender, Dan Flickinger, Stephan Oepen, Woodley Packard, and Ann Copestake. 2015. Layers of interpretation: On grammar and compositionality. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 239–249, London, UK. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Chris Biemann, Kalina Bontcheva, Richard Eckart de Castilho, Iryna Gurevych, and Seid Muhie Yimam. 2017. Collaborative web-based tools for multi-layer text annotation. In Nancy Ide and James Pustejovsky, editors, *The Handbook of Linguistic Annotation*, Text, Speech, and Technology book series, pages 229–256. Springer Netherlands.

P. Blackburn and J. Bos. 2005. *Representation and Inference for Natural Language. A First Course in Computational Semantics*. CSLI.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Geert Booij. 1986. Form and meaning in morphology; the case of dutch agent nouns. *Linguistics*, 24:503–518.

J. Bos. 2003. Implementing the Binding and Accommodation Theory for Anaphora Resolution and Presupposition Projection. *Computational Linguistics*, 29(2):179–210.

Johan Bos. 1996. Predicate Logic Unplugged. In *Proceedings of the Tenth Amsterdam Colloquium*, pages 133–143, ILLC/Dept. of Philosophy, University of Amsterdam.

Johan Bos. 2016. Expressive power of abstract meaning representations. *Computational Linguistics*, 42(3):527–535.

Johan Bos, Valerio Basile, Kilian Evang, Noortje Venhuizen, and Johannes Bjerva. 2017. The Groningen Meaning Bank. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, volume 2, pages 463–496. Springer.

Johan Bos, Kilian Evang, and Malvina Nissim. 2012. Annotating semantic roles in a lexicalised grammar environment. In *Proceedings of the Eighth Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-8)*, pages 9–12, Pisa, Italy.

Johan Bos and Malvina Nissim. 2015. Uncovering noun-noun compound relations by gamification. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*.

Johan Bos and Jennifer Spenader. 2011. An annotated corpus for the analysis of vp ellipsis. *Language Resources and Evaluation*, 45(4):463–494.

Sabine Buchholz and Javier Latorre. 2011. Crowdsourcing preference tests, and how to detect cheating. In *INTERSPEECH-2011*, pages 3053–3056.

Alistair Butler and Kei Yoshimoto. 2012. Banking meaning representations from treebanks. *Linguistic Issues in Language Technology*, 7(6):1–22.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Gregory Norman Carlson. 1977. *Reference to Kinds in English*. Ph.D. thesis, University of Massachusetts.

Richard Eckart de Castilho, Giulia Dore, Thomas Margoni, Penny Labropoulou, and Iryna Gurevych. 2018. A legal perspective on training models for natural language processing. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

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

Stephen Clark and James R. Curran. 2004. 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.

Ann Copestake, Dan Flickinger, Ivan Sag, and Carl Pollard. 2005. Minimal recursion semantics: An introduction. *Journal of Research on Language and Computation*, 3(2–3):281–332.

Mary Dalrymple, Stuart M. Shieber, and Fernando C.N. Pereira. 1991. Ellipsis and Higher-Order Unification. *Linguistics and Philosophy*, 14:399–452.

Donald Davidson. 1967. The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Lucia Donatelli, Michael Regan, William Croft, and Nathan Schneider. 2018. Annotation of tense and aspect semantics for sentential AMR. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 96–108, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rebecca Dridan and Stephan Oepen. 2011. Parser evaluation using elementary dependency matching. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 225–230, Dublin, Ireland. Association for Computational Linguistics.

Kilian Evang, Valerio Basile, Grzegorz Chrupała, and Johan Bos. 2013. Elephant: Sequence labeling for word and sentence segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1426.

Kilian Evang and Johan Bos. 2016. Cross-lingual learning of an open-domain semantic parser. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 579–588, Osaka, Japan.

Christiane Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28.

Dan Flickinger. 2011. Accuracy vs. robustness in grammar engineering. In Emily M. Bender and Jennifer E. Arnold, editors, *Language from a cognitive perspective: Grammar, usage, and processing*, pages 31–50. CSLI Publications.

Daniel Flickinger, Yi Zhang, and Valia Kordoni. 2012. Deepbank: A dynamically annotated treebank of the wall street journal. In *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories*, pages 85–96. Edições Colibri.

Karën Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Last words: Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Annemarie Friedrich and Manfred Pinkal. 2015. Discourse-sensitive automatic identification of generic expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1272–1281, Beijing, China. Association for Computational Linguistics.

Jan Hajič, Eva Hajičová, Marie Mikulová, and Jiří Mírovský. 2017. *Prague Dependency Treebank*. Springer Verlag, Berlin, Germany.

Jerry R. Hobbs. 1991. SRI international's TACITUS system: MUC-3 test results and analysis. In *Proceedings of the 3rd Conference on Message Understanding, MUC 1991, San Diego, California, USA, May 21-23, 1991*, pages 105–107.

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

Nancy Ide and Laurent Romary. 2006. Representing linguistic corpora and their annotations. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

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

Hans Kamp and Antje Rossdeutscher. 2009. Drs construction and lexically driven inferences. *Theoretical Linguistics*, 20(2-3):165–236.

Gene Louis Kim and Lenhart Schubert. 2019. A type-coherent, expressive representation as an initial step to language understanding. *CoRR*.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40.

Martijn van der Klis, Bert Le Bruyn, and Henriëtte de Swart. 2017. Mapping the perfect via translation mining. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 497–502, Valencia, Spain. Association for Computational Linguistics.

Oier Lopez de Lacalle and Eneko Agirre. 2015. Crowdsourced word sense annotations and difficult words and examples. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 94–100, London, UK. Association for Computational Linguistics.

Jochen L. Leidner. 2008. *Toponym Resolution in Text : Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Universal Press, Boca Raton, FL, USA.

Douglas Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38:33–38.

Mike Lewis and Mark Steedman. 2014. A* ccg parsing with a supertag-factored model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 990–1000, Doha, Qatar. Association for Computational Linguistics.

Qianchu Liu, Federico Fancellu, and Bonnie Webber. 2018. NegPar: A parallel corpus annotated for negation. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Emar Maier. 2014. Pure quotation. *Philosophy Compass*, 9(9):615–630.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christopher D. Manning. 2004. Lingo redwoods. *Research on Language and Computation*, 2(4):575–596.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajič, Angelina Ivanova, and Zdeňka Urešová. 2016. Towards comparability of linguistic graph banks for semantic parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 3991–3995, Portorož, Slovenia.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Terence Parsons. 1990. *Events in the semantics of English: A study in subatomic semantics.* Cambridge, MA: The MIT Press.

M.C. Postma, F. Ilievski, and P.T.J.M. Vossen. 2018. Semeval-2018 task 5: Counting events and participants in the long tail. In *SemEval-2018 : International Workshop on Semantic Evaluation 2018*.

M.C. Postma, R. Izquierdo, E. Agirre, G. Rigau, and P.T.J.M. Vossen. 2016. Addressing the mfs bias in wsd systems. In *LREC 2016*, pages 1695–1700.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning - a Guide to Corpus-Building for Applications*. O'Reilly.

Aarne Ranta. 2011. *Grammatical Framework: Programming with Multilingual Grammars*. Center for the Study of Language and Information/SRI.

Nils Reiter and Anette Frank. 2010. Identifying generic noun phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 40–49, Uppsala, Sweden. Association for Computational Linguistics.

Uwe Reyle. 1993. Dealing with Ambiguities by Underspecification: Construction, Representation and Deduction. *Journal of Semantics*, 10:123–179.

Uwe Reyle. 1995. On Reasoning with Ambiguities. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–8, Dublin, Ireland.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Naomi Saphra and Adam Lopez. 2015. AMRICA: an AMR inspector for cross-language alignments. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 36–40, Denver, Colorado. Association for Computational Linguistics.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia. Association for Computational Linguistics.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the eu bookshop corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1850–1855, Reykjavik, Iceland. European Language Resources Association (ELRA).

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

A. Tarski and R. Vaught. 1956. Arithmetical extensions of relational systems. *Compositio Mathematica*, 13:81–102.

David R. Traum. 2000. 20 questions for dialogue act taxonomies. *Journal of Semantics*, 17(1):7–30.

Rob Matthijs van der Goot. 2019. *Normalization and parsing algorithms for uncertain input*. Ph.D. thesis, University of Groningen.

Rik Van Noord, Lasha Abzianidze, Hessel Haagsma, and Johan Bos. 2018. Evaluating scoped meaning representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1685–1693, Miyazaki, Japan.

Noortje Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. 2013. Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, pages 397–403, Potsdam, Germany.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, Texas. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*.