Layers of Interpretation: On Grammar and Compositionality

Emily M. Bender
University of Washington
ebender@uw.edu

Dan Flickinger
Stanford University
danf@stanford.edu

Stephan Oepen
University of Oslo and Potsdam University
oe@ifi.uio.no

Woodley Packard
University of Washington
sweaglesw@sweaglesw.org

Ann Copestake
University of Cambridge
aac@cl.cam.ac.uk

Abstract

With the recent resurgence of interest in semantic annotation of corpora for improved *semantic parsing*, we observe a tendency which we view as ill-advised, to conflate sentence meaning and speaker meaning into a single mapping, whether done by annotators or by a parser. We argue instead for the more traditional hypothesis that sentence meaning, but not speaker meaning, is compositional, and accordingly that NLP systems would benefit from reusable, automatically derivable, task-independent semantic representations which target sentence meaning, in order to capture exactly the information in the linguistic signal itself. We further argue that compositional construction of such sentence meaning representations affords better consistency, more comprehensiveness, greater scalability, and less duplication of effort for each new NLP application. For concreteness, we describe one well-tested grammar-based method for producing sentence meaning representations which is efficient for annotators, and which exhibits many of the above benefits. We then report on a small inter-annotator agreement study to quantify the consistency of semantic representations produced via this grammar-based method.

1 Introduction

Kate and Wong (2010) define 'semantic parsing' as "the task of mapping natural language sentences into complete formal meaning representations which a computer can execute for some domain-specific application." At this level of generality, semantic parsing has been a cornerstone of NLU from its early days, including work seeking to support dialogue systems, database interfaces, or machine translation (Woods et al., 1972; Gawron et al., 1982; Alshawi, 1992, inter alios). What distinguishes most current work in semantic parsing from such earlier landmarks of old-school NLU is (a) the use of (highly) task- and domain-specific meaning representations (e.g. the RoboCup or GeoQuery formal language) and (b) a lack of emphasis on natural language syntax, i.e. a tacit expectation to map (more or less) directly from a linguistic surface form to an abstract representation of its meaning.

This approach risks conflating a distinction that has long played an important role in the philosophy of language and theoretical linguistics (Quine, 1960; Grice, 1968), viz. the contrast between those aspects of meaning that are determined by the linguistic signal alone (called 'timeless', 'conventional', 'standing', or 'sentence' meaning), on the one hand, and aspects of meaning that are particular to a context of use ('utterer', 'speaker', or 'occasion' meaning, or 'interpretation'), on the other hand. Relating this tradition to computational linguistics, Nerbonne (1994, p. 134) notes:

> Linguistic semantics does not furnish a characterization of the *interpretation* of utterances in use, which is what one finally needs for natural language understanding applications—rather, it (mostly) provides a characterization of *conventional content*, that part of meaning determined by linguistic form. Interpretation is not determined by

Proceedings of the 11th International Conference on Computational Semantics, pages 239–249,
London, UK, April 15-17 2015.

> form, however, nor by its derivative content. In order to interpret correctly, one must exploit further knowledge sources and processes that ... probably are not linguistic at all: domain knowledge, common sense, communicative purpose, extralinguistic tasks, assumptions of interlocutors about each other.

In the currently widespread approach to semantic parsing, the results of linguistic research in semantics are largely disregarded in favor of learning correlations between domain-typical linguistic forms and task-specific meaning representations, using the linguistic signal as well as domain-specific information (e.g. a database schema) as sources of constraint on the search space for the machine action that is most likely the one desired. We see two interrelated drawbacks to such an approach: First, to the extent that grammatical structure is taken into account, the same problems must be solved anew with each new task. Second, as a result, such task-specific solutions seem unlikely to scale to general-purpose natural language understanding. In order to reach that lofty goal, we argue, there must be some task-independent model of the conventional content of linguistic utterance types which can be paired with domain-specific knowledge and reasoning in order to reach appropriate interpretations of utterances in context.

Success in many semantically-sensitive NLP tasks requires algorithms that can glean a representation of at least a subset of speaker meaning. But what machines have access to is not any direct representation of a human interlocutor's intended speaker meaning, but rather only natural language utterances. Such utterances involve tokens of sentence (or sentence fragment) types, which in turn have computable sentence meaning. While sentence meaning does not determine situated speaker meaning, it is an important cue to it (Quine, 1960; Grice, 1968; Reddy, 1979; Clark, 1996). We argue here that sentence meaning, but not speaker meaning, is compositional (see Grice 1967), and accordingly that NLP systems would benefit from reusable, automatically derivable, task-independent semantic representations which target sentence meaning, in order to capture exactly the information in the linguistic signal itself. Furthermore, we argue that such sentence meaning representations are best built compositionally, because the compositional approach affords better consistency, more comprehensiveness, and greater scalability.

In this position paper we begin by providing a working definition of compositionality and briefly survey different types of semantic annotation with an eye towards classifying them as compositional or not (§2). §3 provides an overview of the English Resource Grammar (ERG; Flickinger 2000, 2011), a resource for producing semantic representations, covering most of what falls within our definition of compositional, at scale. §4 articulates the three main benefits of a compositional approach to producing semantic annotations, viz. comprehensiveness, consistency and scalability. In §5, we present a small inter-annotator agreement study to quantify the consistency of semantic representations produced via grammar-based sembanking. Finally, in §6, we consider how ERG-based semantic representations can be used as the backbone of even richer annotations that incorporate information which is not compositionally derivable.

2 Compositionality

In this section we explore which aspects of meaning among those captured by annotation projects serving NLP work (and thus presumably of interest to the NLP community) can be seen as compositional. Our purpose in doing so is two-fold: On the one hand, it illuminates the claim that sentence meaning is compositional by delineating those aspects of meaning representations admissible as sentence meaning by that criterion. On the other hand, it sheds light on the range of possible contributions of a grammar-based approach to semantic annotation.

As Szabó (2013) points out, there are many different interpretations of the principle of compositionality in the literature. Since we are concerned with annotation, the issue is compositionality of meaning representations (rather than denotation, for instance). In order to ask which aspects of meaning are compositional, we provide the following working definition:[1]

[1] In Szabó's terms, our definition of compositionality is local, distributive, and language-bound and furthermore consistent with the rule-to-rule principle. It is also consistent with the notion of compositionality from Copestake et al. (2001) and implemented in the ERG, which furthermore adds the constraint that the function for determining meanings of complex expressions must be monotonic in the sense that it cannot remove or overwrite any information contributed by the constituents.

(1) A meaning system (or subsystem) is compositional if:

- it consists of a finite (but possibly very large) number of arbitrary atomic symbol-meaning pairings;
- it is possible to create larger symbol-meaning pairings by combining the atomic pairings through a finite set of rules;
- the meaning of any non-atomic symbol-meaning pairing is a function of its parts and the way they are combined;
- this function is possibly complex, containing special cases for special types of syntactic combination, but only draws on the immediate constituents and any semantic contribution of the rule combining them; and
- further processing will not need to destructively change a meaning representation created in this way to create another of the same type.

Applying this definition to layers of semantic annotation attested in various projects within NLP, we find that they include both compositional and non-compositional aspects of meaning.

Perhaps the clearest candidate for an annotation layer that is compositional is predicate-argument structure, which appears to be fully grammar-derived: Lexical entries (atoms) provide predicates and argument positions; grammar rules dictate the linking of arguments across predicates. Note, however, that there may be disagreement as to whether particular linkings (e.g. between the subject of a participial modifier and the subject of the clause it modifies) are required by the grammar or simply anaphoric in nature. Beyond predicate-argument structure, the grammars of particular languages also provide at least partial constraints on the scope of negation and other operators, the restriction of quantifiers, modality, tense/aspect/mood, information structure, discourse status of referents of NPs, and politeness. These subsystems we consider partially compositional. There are also layers of annotation that may be considered compositional, but not according to sentence grammar, such as coherence relations/rhetorical structure.

Turning to types of semantic annotation which are not compositional, we first find layers that concern only atoms. These include fine-grained word-sense tagging, named entity tags and so on. According to the definition we have given, there may be an indefinite number of atom-meaning pairings, but these are outside the scope of the compositionality principle. What is built compositionally, on our account, is the relationships between the pieces of meaning contributed by the words.[2] There is an additional principle, often tacitly assumed, that word-meaning pairings should not be multiplied beyond necessity: in the strictest form of this, word senses are only distinguished if the distinction interacts with the syntax and morphology. A compositional representation that is consistent with this principle can be further specialized with finer-grained word sense and semantic role information without changing its structure, and hence this amounts to a form of underspecification, rather than a strong claim about lexical meaning.

Another kind of non-compositional meaning layer is that which requires some sort of further computation over linguistic structure. This can be seen as purely monotonic addition of further constraints on underspecified meaning representations, but it is not compositional in the sense that it is never (strictly) constrained by grammatical structure. In this category, we find quantifier scope ambiguity resolution (e.g. Higgins and Sadock 2003), coreference resolution (e.g. Hobbs 1979), and the determination of the focus of negation (e.g. Blanco and Moldovan 2011). All of these build on partial constraints provided by the grammar, but in all cases, the interpretation of particular sentences in context will correspond to one (or a subset) of the possibilities allowed by the grammar.

The next layer of meaning annotation to consider corresponds to discourse processing. This includes the calculation of presupposition projection (e.g. Van der Sandt 1992; Zaenen and Karttunen 2013; Venhuizen et al. 2013), coherence relations/rhetorical structure (e.g. Marcu 1997), and the annotation of discourse moves/adjacency pairs (Shriberg et al., 2004). These aspects of meaning clearly build on information provided during sentence-level processing, including lexically determined veridicality contexts (e.g. (2a) vs. (2b)) as well as discourse connectives. In both cases, the grammatical structure links embedded clauses to the relevant lexical predicates.

[2] We assume here that word sense is a property of roots, rather than fully inflected forms. Productive derivational morphology supports compositionally built-up meanings for morphologically complex words. Semi-productive morphological processes and frozen or lexicalized complex forms complicate any conventional grammar-based treatment, however.

(2) a. They forgot to vote. [⇒ They didn't vote.]

b. They forgot that they had voted. [⇒ They did vote.]

As this level of processing concerns relationships both within and across sentences, it is clearly not compositional with respect to sentence grammar. We consider it an open question whether there are compositional processes at higher levels of structure that constrain these aspects of meaning in analogous ways, but we note that in presupposition processing at least, a notion of defeasibility is required (Asher and Lascarides, 2011).

Finally, there are semantic annotations that attempt to capture what speakers are trying to do with their speech acts. This includes tasks like hedge detection (Vincze et al., 2008) and the annotation of social acts such as authority claims and alignment moves (Morgan et al., 2013) or the pursuit of power in dialogue (Swayamdipta and Rambow, 2012). While in some cases there are keywords that have a strong association with particular categories in these annotation schemes, these aspects of meaning are clearly not anchored in the structure of sentences but rather relate to the goals that speakers have in uttering sentences. Lacking a firm link to the structure of sentences, they do not appear to be compositional.

We have seen in this (necessarily brief) section that existing annotation projects span both compositional and non-compositional aspects of meaning, and we have furthermore identified those that are constrained, in whole or in part, by (sentence-level) grammatical structures as well as those that build on such structures. In the following section we describe how the English Resource Grammar relates to the broader project of creating rich representations of sentence meaning.

3 The English Resource Grammar

We have categorized layers of meaning annotation according to whether they are compositional, and if not, in what way they fail to demonstrate compositionality. The majority of those identified as compositional are compositional according to sentence grammar. In this section, we briefly describe the English Resource Grammar (ERG; Flickinger 2000, 2011) an open-source, domain-independent, linguistically precise, broad-coverage grammar for English that encapsulates the linguistic knowledge required to produce many of the types of compositional meaning annotations described above, at scale.

The ERG is an implementation of the grammatical theory of Head-driven Phrase Structure Grammar (HPSG; Pollard and Sag 1994), i.e. a computational grammar that can be used for both parsing and generation. Development of the ERG started in 1993, building conceptually on earlier work on unification-based grammar engineering at Hewlett Packard Laboratories (Gawron et al., 1982). The ERG has continuously evolved through a series of research projects (and two commercial applications) and today allows the grammatical analysis of most running text across domains and genres. In the most recent stable release, version '1214', the ERG contains 225 syntactic rules and 70 lexical rules for derivation and inflection. The hand-built ERG lexicon of some 39,000 lemmata, instantiating 975 leaf lexical types providing part-of-speech and valence constraints, aims for complete coverage of function words and open-class words with 'non-standard' syntactic properties (e.g. argument structure). Built-in support for light-weight named entity recognition and an unknown word mechanism typically enable the grammar to derive full syntactico-semantic analyses for 85–95% of all utterances in standard corpora, including newspaper text, the English Wikipedia, or bio-medical research literature (Flickinger et al., 2012, 2010; Adolphs et al., 2008). Each of these analyses includes both a derivation tree recording the grammar rules and lexical entries that were used, and the associated semantic representation produced compositionally via this derivation, within the Minimal Recursion Semantics (MRS) framework (Copestake et al., 2005). We refer to these ERG-derived MRS objects stored in Redwoods treebanks as English Resource Semantics (ERS) expressions.

Using the annotation methodology described by Oepen et al. (2004), for each roughly annual release of the ERG, a selection of development corpora is manually annotated with the 'intended' analysis among the alternatives provided by the grammar. For those utterances which either do not receive a full parse or where no correct analysis can be found by the annotator in the available parse forest, partial analyses

can be assigned during annotation by making use of a recursively applicable binary *bridging* rule which preserves the semantic contributions of its two daughter constituents at each application of the rule. Alternatively, the annotator may record that no analysis is available. The automatic exports of the correct derivation tree and ERS for each of these sentences are made available as the Redwoods Treebank; at the end of 2014, the current version of Redwoods encompasses gold-standard ERG analyses for 85,000 utterances (∼1.5 million tokens) of running text from half a dozen different genres and domains.

In more detail, the task of annotation for a sentence consists of making binary decisions about the set of *discriminants* each of which partitions the parse forest into two: all of the analyses which employ the particular rule or lexical entry, and the rest of the analyses which do not. This method, originating with Carter 1997, enables the human annotator to rapidly discard analyses in order to isolate the intended analysis, or to conclude that the correct analysis is unavailable. As a reference point for speed of annotation using this method, an expert treebanker using the current '1214' version of the ERG annotated 2400 sentences (37,200 words) from the Brown corpus in 1400 minutes, for an average rate of 1.7 sentences per minute.[3]

Annotations produced by this method of choosing among the candidate analyses licensed by a grammar will thus record those components of sentence meaning which are constrained by the grammatical structure and lexical items used in the intended analysis. In the next section we review some of the desired benefits of this method for producing and maintaining semantically annotated corpora which are sufficiently detailed, consistent, and numerous to be of use in non-trivial NLP applications that require the computation of semantics either for parsing or for generation.

4 Benefits of Compositionality

We are concerned here with the goal of designing task-independent semantic representations and deploying them at scale to create a large sembank including diverse genres. Such representations can be created compositionally, where the content and internal structure of the representations is constrained by syntactic structure, or non-compositionally, where annotators encode their understanding of a sentence directly. The latter category is exemplified by Abstract Meaning Representation (AMR; Langkilde and Knight 1998; Banarescu et al. 2013). In the former category, we find both manual annotation projects, such as PropBank (Kingsbury and Palmer, 2002) and FrameNet (Baker et al., 1998), which annotate semantic information with reference to syntactic structure, and grammar-based annotation initiatives such as the Redwoods Treebank (Oepen et al., 2004), TREPIL (Rosén et al., 2005), and the Groningen Meaning Bank (Basile et al., 2012).

We argue here that a grammar-based, compositional approach is critical to achieving this long-range goal, in particular because it supports more comprehensive representations (§4.1), produced with better consistency (§4.2) and greater scalability (§4.3). The drawback to a grammar-based approach is that it cannot, in itself, include information that is not compositional, but as we will develop further in §6 below, it is possible to have the best of both worlds, adding non-compositional information as additional annotation layers over grammar-produced semantic representations.

4.1 Comprehensiveness

Where task-specific meaning representations are free to abstract away from task-irrelevant details of linguistic expression, task-independent representations only have that luxury when the variation is truly (sentence) meaning preserving. A task-independent semantic representation should capture exactly the meaning encoded in the linguistic signal itself, as it is is not possible to know, *a priori*, which parts of that sentence meaning will be critical to determining speaker meaning in any given application.

[3]This rate is roughly consistent with an earlier experiment using the same Redwoods treebanking method where annotation times were noted: MacKinlay et al. (2011) report a somewhat slower mean annotation time by an expert annotator of 0.6 sentences per minute, but this difference can be attributed to the greater average sentence length (and hence increased number of discriminants to be determined) for that biomedical corpus: 23.4 tokens compared with 15.5 for the Brown data.

⟨ h_1,
| h_4:person⟨0:6⟩(ARG0 x_5),
| h_6:_no_q⟨0:6⟩(ARG0 x_5, RSTR h_7, BODY h_8),
| h_2:_eat_v_1⟨7:11⟩(ARG0 e_3, ARG1 x_5, ARG2 i_9)
{ $h_1 =_q h_2$, $h_7 =_q h_4$ } ⟩

⟨ h_1,
| h_4:_every_q⟨0:5⟩(ARG0 x_6, RSTR h_7, BODY h_5),
| h_8:_person_n_1⟨6:12⟩(ARG0 x_6),
| h_2:_fail_v_1⟨13:19⟩(ARG0 e_3, ARG1 h_9),
| h_{10}:_eat_v_1⟨23:27⟩(ARG0 e_{11}, ARG1 x_6, ARG2 i_{12})
{ $h_1 =_q h_2$, $h_7 =_q h_8$, $h_9 =_q h_{10}$ } ⟩

```
(e / eat-01
 :polarity -
 :ARG0 (p / person
   :mod (e / every)))
```

Figure 1: ERS and AMR representations of (3a,b).

A grammar-based, compositional approach requires that each word and syntactic structure in a sentence be accounted for, either by contributing its piece to the overall semantic representation or by being explicitly determined to be semantically vacuous.[4] As a result, the paraphrase sets in a grammar-based approach tend to be narrower than those produced by non-compositional, free-hand annotation. For example, the AMR annotation guidelines (Banarescu et al., 2014) present the examples in (3) and (4) as paraphrase sets, as they are likely be true in many of the same real-world circumstances and/or have the same practical interpretation in a given task.

(3) a. No one ate.
    b. Every person failed to eat.

(4) a. The boy is responsible for the work.
    b. The boy is responsible for doing the work.
    c. The boy has the responsibility for the work.

Nonetheless, these sentences intuitively differ in nuances of meaning. Compare these to (5), which gives a (partial) set of strings analyzed as exact paraphrases by the ERG:

(5) a. Kim thinks Sandy gave the book to Pat.
    b. Kim thinks that Sandy gave the book to Pat.
    c. Kim thinks Sandy gave Pat the book.
    d. Kim thinks the book was given to Pat by Sandy.
    e. The book, Kim thinks Sandy gave to Pat.

The members of the paraphrase sets identified by the ERG, while varying in interesting ways in their syntax, remain very close in their lexical content and share semantic dependencies. In contrast, the ERG assigns distinct representations to the different items in (3) and (4), as shown in Figure 1.

To summarize, compositionally constructed meaning representations are comprehensive in the sense that they account for every meaningful component of the string. We note that natural languages feature meaningful elements whose contribution is orthogonal to questions of truth conditions, for example, politeness markers (e.g. the word *please* in English, or grammaticized elements such as pronoun choice in many other languages) and markers of speaker attitude towards what is being expressed (e.g. adverbs like *frankly*). A comprehensive meaning representation should also capture both the contributions of these types of elements and their relations to the meaning of the rest of the sentence.

4.2 Consistency

A compositional approach to semantic annotation promotes consistency in the first instance by imposing constraints on possible semantic representations. Requiring meaning representations to be grounded in both the lexical items and syntactic structure of the strings being annotated significantly reduces the space of possible annotations. Grammar-based compositional approaches add to this the ability to encode design decisions about the annotations in machine-readable form and thus automatically produce only representations that conform to the design decisions. For example, the two arguments of the subordinating conjunction *when* are relatively similar in their relationship to the predicate. A grammar-based approach can with complete consistency always map the clause immediately after *when* to the same one of those arguments.

[4]Determining which elements are semantically vacuous can be non-trivial, but generally turns on testing paraphrase candidates for truth-conditional equivalence.

This does not mean that no human effort is required in grammar-based annotation, but the annotation task is simpler.[5] In grammar-based sembanking, discussed further in §5 below, the annotators are choosing among representations created by the grammar, rather than creating the representations themselves. One way to quantify the relative simplicity of such an approach is in the length of the annotation guidelines. For the study reported in §5, we relied on a set of heuristics explained in <1,000 words, along with documentation of the grammar explaining the purpose of each grammar rule and type of lexical entry. This contrasts with the AMR annotation guidelines (Banarescu et al., 2014), which run to 57 pages, or the PropBank annotation guidelines (Babko-Malaya, 2005), at 38 pages. The more human annotators can rely on their linguistic intuitions (here, in judging whether the grammar-produced semantic representations match their understanding of the most likely intended sentence meaning in context) rather than trying to track and apply a wide range of rules or heuristics, the more we can expect them to produce consistent results. A second way in which a compositional approach promotes consistency is by providing 'guide rails' to help annotators adhere to sentence meaning as opposed to speaker meaning. In human annotation of linguistic data, annotators will always be working in the context of their own interpretation of the intended speaker meaning. A sentence-meaning annotation task requires annotators to separate out their intuitions about the one from the other. If the target annotations are grammatically constrained (with or without the aid of a computational grammar), it should be more feasible for annotators to restrict themselves to annotating those parts of the meaning which are due to the linguistic signal itself.

4.3 Scalability

A third benefit of a compositional approach is that, by enabling grammar-based annotation, it gains great scalability, in terms of the amount of text annotated, genre diversity and long-term refinement of the annotations. While the initial effort required to create an implemented grammar exceeds the (still not inconsiderable) effort required to create and pilot annotation guidelines, once an implemented grammar has reached an interesting level of coverage, it can be used to annotate text very quickly. This makes it inexpensive to incorporate new genres into the collection of annotated text. Flickinger (2011) observes that the ERG has fairly consistent coverage across quite divergent genres (including tourist brochures, Wikipedia articles, online user forum posts, and chemistry papers) since the higher-frequency, core phenomena are the same.

However, even the most complete linguistically precise grammar will still lack complete coverage over naturally occurring texts, if only because of ungrammatical or extragrammatical strings in those texts. This is the familiar trade-off between accuracy (or, more relevantly here, consistency) and robustness. However, for many applications, complete annotation of the input text is not required, but merely a sufficiently large sample of annotated in-domain items. Furthermore, a precision grammar can be augmented with robustness strategies, at a cost in annotation detail and/or consistency. This 'bridging'-rule approach to this problem is discussed further in §5 below.

Finally we observe that grammar-based semantic annotation also scales particularly well in the complexity of the annotations themselves. Specifically, discriminant-based treebanking/sembanking (Carter, 1997) supports a dynamic approach to annotation that allows the annotated resources to be updated largely automatically when the grammar underlying the annotations is improved (Oepen et al., 2004). In the present context, this means that any refinements to the analysis of particular semantic phenomena or additions of layers of grammar-based annotation (e.g. information structure) that are added to the grammar can be swiftly applied throughout the corpus.

4.4 Summary

In this section we have elaborated on what we consider the three main benefits to a compositional approach to semantic annotation: comprehensiveness, consistency and scalability. In the following section, we provide a quantitative study of consistency as well as some quantitative indicators of scalability.

[5] The development of a grammar in the first place represents a lot of human effort, but this effort is captured in an artifact—the grammar—that allows it to be reused within and across domains indefinitely; see §4.3.

5 Inter-Annotator Agreement in Grammar-Based Sembanking

We carried out a small-scale experiment to quantify the consistency achievable with grammar-based sembanking specifically using the English Resource Grammar. For comparability with results reported by the AMR project (Banarescu et al., 2013), we drew our text from Antoine de Saint-Exupéry's *The Little Prince*. The annotations were produced by three of the authors, according to the Redwoods discriminant-based treebanking methodology (see §3 above), with some extensions (discussed below). We first triply annotated a 50-sentence trial set and then produced an adjudicated gold standard for that set, refining and documenting our annotation heuristics in the process. We then proceeded to independently annotate a 150-sentence sample. It is on this larger sample that we report inter-annotator agreement.

In our study, we used revised annotation software, allowing two key improvements over the classical Redwoods procedure: The first of these enhancements makes unbiased annotation possible even in cases where the level of ambiguity makes complete enumeration of the space of candidate analyses computationally prohibitive. In the original Redwoods set-up, annotators are only presented with the top N (usually 500) analyses, according to a parse selection model trained on previously annotated data. In the revised approach (Packard, 2015), we compute discriminants directly from a packed parse chart which preserves all of the parses in the forest, rather than by comparing a subset of the individual analyses to each other. Second, where previously Redwoods-style treebanking was limited to those sentences for which the grammar could produce a correct analysis, we have adopted a technique that allows us to produce meaning representations for all sentences in the input—though these representations will be incomplete in cases where the grammar does not find a correct spanning analysis. The technique involves augmenting the grammar with two pseudo-grammatical rules, one which projects any grammatical constituent to an 'island', and one which bridges two adjacent islands. The semantic representations within each island will be consistent with the standards of the grammar, though the connections between islands are left vague. Since this extension is not intended to produce additional analyses for items which are correctly analysed, sembanking is done in two passes: first, with the robust analyses suppressed, and then on a second pass, only for items without satisfactory analyses, with the bridging rules turned on. In our sample sembank, these robust analyses allowed us to increase the coverage from between 79% and 88% of sentences (depending on the annotator) to 100% (for all three annotators).

We also produced an adjudicated gold standard version of all 200 annotated sentences.[6] This was achieved by comparing the annotations selected by each annotator (with or without bridges), for each item on which there was disagreement (71 items, including 55 of the 150 item sample), discussing the differences, and either selecting one of the three as fully correct or creating a hybrid representing the consensus decision for each choice point. When we felt that the decisions were not already fully guided by the existing annotation guidelines, we worked to articulate an extension to the guidelines that would support the decision.

Table 1 summarizes inter-annotator agreement for the 150 item sample, using three different metrics:[7] (a) *exact match*, i.e. wholly identical ERSs, and F_1 scores for (b) *argument identification* (EDM_a) and (c) for *predicate naming and argument identification* (EDM_{na}). The latter two metrics quantify what Dridan and Oepen (2011) call Elementary Dependency Match, i.e. F_1 computed over sets of triples extracted from a reduction of each full ERS into a variable-free Elementary Dependency Structure (EDS; Oepen and Lønning 2006), i.e. a labeled, directed graph that is formally very similar to AMRs. Here, we consider two types of triples, viz. ones associating a predicate name with a node and ones relating two nodes in an argument relation. From the leftmost ERS in Figure 1, for example, three predicate name triples would be extracted, including $\langle 7{:}11\rangle$ – NAME – _eat_v_1, and three argument triples (discarding unexpressed arguments), including $\langle 7{:}11\rangle$ – ARG1 – $\langle 0{:}6\rangle$.[8] Slightly higher EDM_{na} than EDM_a suggests that predicate naming is easier to annotate than argument identification, which is plausible seeing that

[6] The adjudicated gold standard, together with the annotation guidelines, is available from `www.delph-in.net/lpp`.

[7] None of these correct for chance agreement as that is currently an unsolved methodological problem in graph-structured annotations

[8] For increased comparability with AMR, we ignore a third type of triple, corresponding to so-called variable properties, i.e. information about tense, mood, aspect, number, or person.

	Annotator Comparison			
Metric	**A vs. B**	**A vs. C**	**B vs. C**	**Average**
Exact Match	0.73	0.65	0.70	0.70
EDM_a	0.93	0.92	0.94	0.93
EDM_{na}	0.94	0.94	0.95	0.94

Table 1: Exact match ERS and Elementary Dependency Match across three annotators.

the only sense distinctions drawn in the ERG are those that contrast in their syntactic distribution.

In general, the statistics in Table 1 strongly support our expectations regarding consistency of annotation: In at least two out of three cases, any pair of annotators arrives at exactly the same representation of sentence meaning; in the granular EDM metrics, pairwise inter-annotator agreement ranges consistently in the mid-nineties F_1. Although not directly comparable due to methodological differences in the interpretation of the task of sembanking, formal differences in the nature of target representations, and the annotation of web texts rather than *The Little Prince*, we observe that Banarescu et al. (2013) report triple-based F_1 scores for inter-annotator agreement in AMR sembanking of 0.71. This, in our view, clearly suggests that a well-defined target representation at the level of sentence meaning (or, in other words, all and only the grammaticized contributions to interpretation) affords comparatively high levels of annotation quality at relatively moderate costs per additional items annotated.

6 Conclusion: Further Layers of Annotation

In this position paper, we have argued that NLP systems would benefit from task-independent semantic representations which capture the information in the linguistic signal (i.e. sentence meaning), as a basis from which to map to task-dependent hypotheses about speaker intentions. Some of the information we would like to see in such annotations is grammatically constrained, and we have argued that representations of those aspects of meaning are best built compositionally. However, there are further aspects of meaning which are closely tied to the linguistic signal but are not constrained by sentence-level grammar (or only partially so constrained). We agree here with Basile et al. (2012) and Banarescu et al. (2013) that a single resource that combines multiple different types of semantic annotations, all applied to the same text, will be most valuable (see also Ide and Suderman 2007). However, just because some aspects of the desired representations cannot be created in a grammar-based fashion does not mean that what can be done with a grammar has no value. To get the best of both worlds, one should start from grammar-derived semantic annotations and then either add further layers of annotation (e.g. word sense, coreference) or, should larger paraphrase sets be desired, systematically simplify aspects of the grammar-derived representations, effectively 'bleaching' some of the contrasts.

In moving from the current state of the art towards more comprehensive representations, we envision several kinds of enrichment: First, there are extensions to the grammar, supporting ever greater coverage and more detailed information about sentence-level, compositional meaning representations (e.g. partial constraints on coreference, reflecting binding-theoretic configurations or partial constraints on information structure). Second, the output of a grammar like the ERG might serve as the input to another sort of grammar to compute structure-sensitive, cross-sentential phenomena such as presupposition propagation. Third, the atoms of the grammar-derived semantic representations could be further annotated with links to word-sense inventories, ontological resources, and the like. Similarly, phenomena such as (non-grammatically constrained) coreference could be annotated over the semantic representations, as links between semantic variables, including those representing syntactically null anaphora.

Critically, this layered approach builds on top of compositional meaning representations. As we have argued, the compositional approach supports the development of more comprehensive representations (capturing more detail at each layer considered), more consistent deployment of the annotations (as design decisions are coded directly into the grammar and the task for the human annotator is simpli-

fied), and greater scalability of annotations, across texts, genres, and in the addition of layers and other refinements to the annotations themselves.

References

Adolphs, P., S. Oepen, U. Callmeier, B. Crysmann, D. Flickinger, and B. Kiefer (2008). Some fine points of hybrid natural language parsing. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Alshawi, H. (Ed.) (1992). *The Core Language Engine*. Cambridge, MA, USA: MIT Press.

Asher, N. and A. Lascarides (2011). Reasoning dynamically about what one says. *Synthese 183*(1), 5 – 31.

Babko-Malaya, O. (2005). PropBank annotation guidelines. Available from `http://verbs.colorado.edu/~mpalmer/projects/ace/PBguidelines.pdf`.

Baker, C. F., C. J. Fillmore, and J. B. Lowe (1998). The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, pp. 86 – 90.

Banarescu, L., C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, Sofia, Bulgaria, pp. 178 – 186.

Banarescu, L., C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider (2014). Abstract Meaning Representation (AMR) 1.1 specification. Version of February 11, 2014.

Basile, V., J. Bos, K. Evang, and N. Venhuizen (2012). Developing a large semantically annotated corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey, pp. 3196 – 3200.

Blanco, E. and D. Moldovan (2011). Semantic representation of negation using focus detection. In *Proceedings of the 49th Meeting of the Association for Computational Linguistics*, Portland, OR, USA, pp. 581 – 589.

Carter, D. (1997). The TreeBanker. A tool for supervised training of parsed corpora. In *Proceedings of the Workshop on Computational Environments for Grammar Development and Linguistic Engineering*, Madrid, Spain, pp. 9 – 15.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Copestake, A., D. Flickinger, C. Pollard, and I. A. Sag (2005). Minimal Recursion Semantics. An introduction. *Research on Language and Computation 3*(4), 281 – 332.

Copestake, A., A. Lascarides, and D. Flickinger (2001). An algebra for semantic construction in constraint-based grammars. In *Proceedings of the 39th Meeting of the Association for Computational Linguistics*, Toulouse, France, pp. 140 – 147.

Dridan, R. and S. Oepen (2011). Parser evaluation using elementary dependency matching. In *Proceedings of the 12th International Conference on Parsing Technologies*, Dublin, Ireland, pp. 225 – 230.

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

Flickinger, D. (2011). Accuracy vs. robustness in grammar engineering. In E. M. Bender and J. E. Arnold (Eds.), *Language from a Cognitive Perspective: Grammar, Usage, and Processing*, pp. 31–50. Stanford: CSLI Publications.

Flickinger, D., S. Oepen, and G. Ytrestøl (2010). WikiWoods. Syntacto-semantic annotation for English Wikipedia. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valletta, Malta.

Flickinger, D., Y. Zhang, and V. Kordoni (2012). DeepBank. A dynamically annotated treebank of the Wall Street Journal. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories*, Lisbon, Portugal, pp. 85 – 96. Edições Colibri.

Gawron, J. M., J. King, J. Lamping, E. Loebner, E. A. Paulson, G. K. Pullum, I. A. Sag, and T. Wasow (1982). Processing English with a Generalized Phrase Structure Grammar. In *Proceedings of the 20th Meeting of the Association for Computational Linguistics*, Toronto, Ontario, Canada, pp. 74 – 81.

Grice, H. P. (1967). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and Semantics 3: Speech Acts*, pp. 41 – 58. New York: Academic Press.

Grice, H. P. (1968). Utterer's meaning, sentence-meaning, and word-meaning. *Foundations of Language 4*(3), 225 – 242.

Higgins, D. and J. M. Sadock (2003). A machine learning approach to modeling scope preferences. *Computational Linguistics 29*(1), 73 – 96.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science 3*(1), 67 – 90.

Ide, N. and K. Suderman (2007). GrAF: A Graph-based Format for Linguistic Annotations. In *Proceedings of the Linguistic Annotation Workshop*, pp. 1 – 8.

Kate, R. J. and Y. W. Wong (2010). Semantic parsing. The task, the state of the art and the future. In *Tutorial Abstracts of the 20th Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp. 6.

Kingsbury, P. and M. Palmer (2002). From TreeBank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain, pp. 1989 – 1993.

Langkilde, I. and K. Knight (1998). Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Meeting of the Association for Computational Linguistics*, Montréal, Canada, pp. 704 – 710.

MacKinlay, A., R. Dridan, D. Flickinger, S. Oepen, and T. Baldwin (2011). Using external treebanks to filter parse forests for parse selection and treebanking. In *Proceedings of the 2011 International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, pp. 246 – 254.

Marcu, D. (1997). The rhetorical parsing of unrestricted natural language texts. In *Proceedings of the 35th Meeting of the Association for Computational Linguistics and the 7th Meeting of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain, pp. 96 – 103.

Morgan, J. T., M. Oxley, E. M. Bender, L. Zhu, V. Gracheva, and M. Zachry (2013). Are we there yet? The development of a corpus annotated for social acts in multilingual online discourse. *Dialogue & Discourse 4*, 1 – 33.

Nerbonne, J. (1994). Book review. Computational linguistics and formal semantics. *Computational Linguistics 20*(1), 131 – 136.

Oepen, S., D. Flickinger, K. Toutanova, and C. D. Manning (2004). LinGO Redwoods. A rich and dynamic treebank for HPSG. *Research on Language and Computation 2*(4), 575 – 596.

Oepen, S. and J. T. Lønning (2006). Discriminant-based MRS banking. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, pp. 1250 – 1255.

Packard, W. (2015). Full forest treebanking. Master's thesis, University of Washington.

Pollard, C. and I. A. Sag (1994). *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. Chicago, USA: The University of Chicago Press.

Quine, W. V. O. (1960). *Word and Object*. Cambridge, MA, USA: MIT Press.

Reddy, M. J. (1979). The conduit metaphor. A case of frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and Thought*, pp. 164 – 201. Cambridge, UK: Cambridge University Press.

Rosén, V., P. Meurer, and K. De Smedt (2005). Constructing a parsed corpus with a large LFG grammar. In M. Butt and T. H. King (Eds.), *Proceedings of the 10th International LFG Conference*, Bergen, Norway.

Shriberg, E., R. Dhillon, S. Bhagat, J. Ang, and H. Carvey (2004). The icsi meeting recorder dialog act (mrda) corpus. In M. Strube and C. Sidner (Eds.), *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, Cambridge, Massachusetts, USA, pp. 97 – 100.

Swayamdipta, S. and O. Rambow (2012). The pursuit of power and its manifestation in written dialog. In *ICSC*, pp. 22 – 29.

Szabó, Z. G. (2013). Compositionality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2013 ed.).

Van der Sandt, R. A. (1992). Presupposition projection as anaphora resolution. *Journal of semantics 9*(4), 333 – 377.

Venhuizen, N., J. Bos, and H. Brouwer (2013). Parsimonious semantic representations with projection pointers. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, Potsdam, Germany, pp. 252–263.

Vincze, V., G. Szarvas, R. Farkas, G. Móra, and J. Csirik (2008). The BioScope corpus. Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics 9(Suppl 11)*.

Woods, W. A., R. M. Kaplan, and B. L. Nash-Webber (1972). The lunar sciences natural language information system. Final report. BBN Report 2378, Bolt Beranek and Newman Inc., Cambridge, MA, USA.

Zaenen, A. and L. Karttunen (2013). Veridicity annotation in the lexicon? A look at factive adjectives. In *Proceedings of the 9th Joint ISO–ACL SIGSEM Workshop on Interoperable Semantic Annotation*, Potsdam, Germany, pp. 51 – 58.