# Evaluating Scoped Meaning Representations

**Rik van Noord, Lasha Abzianidze, Hessel Haagsma, Johan Bos**
CLCG, University of Groningen
{r.i.k.van.noord, l.abzianidze, hessel.haagsma, johan.bos}@rug.nl

**Abstract**

Semantic parsing offers many opportunities to improve natural language understanding. We present a semantically annotated parallel corpus for English, German, Italian, and Dutch where sentences are aligned with scoped meaning representations in order to capture the semantics of negation, modals, quantification, and presupposition triggers. The semantic formalism is based on Discourse Representation Theory, but concepts are represented by WordNet synsets and thematic roles by VerbNet relations. Translating scoped meaning representations to sets of clauses enables us to compare them for the purpose of semantic parser evaluation and checking translations. This is done by computing precision and recall on matching clauses, in a similar way as is done for Abstract Meaning Representations. We show that our matching tool for evaluating scoped meaning representations is both accurate and efficient. Applying this matching tool to three baseline semantic parsers yields F-scores between 43% and 54%. A pilot study is performed to automatically find changes in meaning by comparing meaning representations of translations. This comparison turns out to be an additional way of (i) finding annotation mistakes and (ii) finding instances where our semantic analysis needs to be improved.

**Keywords:** parallel corpus, semantic annotation, discourse representation structure, evaluation, semantic scope

## 1 Introduction

Semantic parsing is the task of assigning meaning representations to natural language expressions. Informally speaking, a meaning representation describes *who did what to whom, when, and where, and to what extent this is the case or not*. The availability of open-domain, wide coverage semantic parsers has the potential to add new functionality, such as detecting contradictions, verifying translations, and getting more accurate search results. Current research on open-domain semantic parsing focuses on supervised learning methods, using large semantically annotated corpora as training data.

However, there are not many annotated corpora available. We present a parallel corpus annotated with formal meaning representations for English, Dutch, German, and Italian, and a way to evaluate the quality of machine-generated meaning representations by comparing them to gold standard annotations. Our work shows many similarities with recent annotation and parsing efforts around Abstract Meaning Representations, (AMR; Banarescu et al., 2013) in that we abstract away from syntax, use first-order meaning representations, and use an adapted version of SMATCH (Cai and Knight, 2013) for evaluation. However, we deviate from AMR on several points: meanings are represented by scoped meaning representations (arriving at a more linguistically motivated treatment of modals, negation, presupposition, and quantification), and the non-logical symbols that we use are grounded in WordNet (concepts) and VerbNet (thematic roles), rather than PropBank (Palmer et al., 2005). We also provide a syntactic analysis in the annotated corpus, in order to derive the semantic analyses in a compositional way.

We make the following contributions:

- A meaning representation with explicit scopes that combines WordNet and VerbNet with elements of formal logic (Section 2).
- A gold standard annotated parallel corpus of formal meaning representations for four languages (Section 3).
- A tool that compares two scoped meaning representations for the purpose of evaluation (Section 4 and Section 5).

## 2 Scoped Meaning Representations

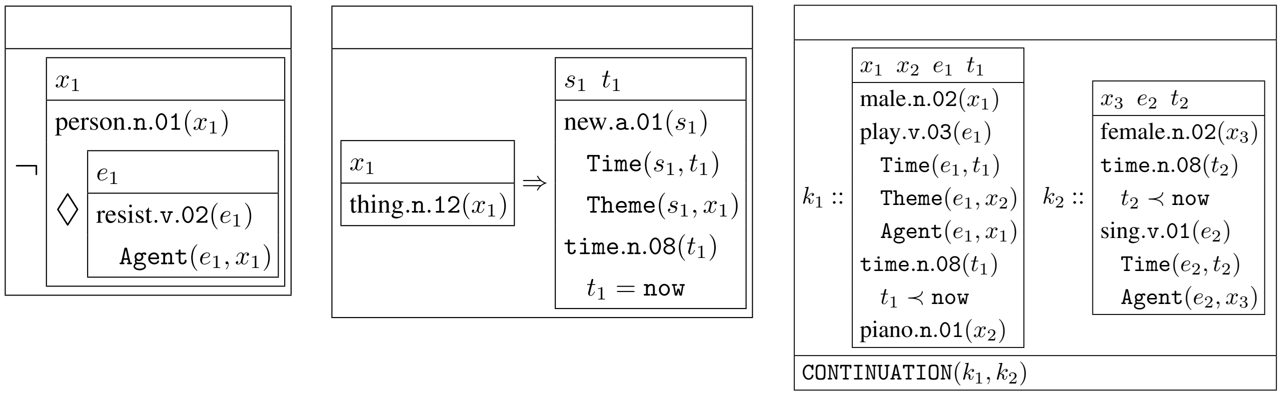
### 2.1 Discourse Representation Structures

The backbone of the meaning representations in our annotated corpus is formed by the Discourse Representation Structures (DRS) of Discourse Representation Theory (Kamp and Reyle, 1993). Our version of DRS integrates WordNet senses (Fellbaum, 1998), adopts a neo-Davidsonian analysis of events employing VerbNet roles (Bonial et al., 2011), and includes an extensive set of comparison operators. More formally, a DRS is an ordered pair of a set of variables (discourse referents) and a set of conditions. There are basic and complex conditions. Terms are either variables or constants, where the latter ones are used to account for indexicals (Bos, 2017). Basic conditions are defined as follows:

- If W is a symbol denoting a WordNet concept and x is a term, then W(x) is a basic condition;
- If V is a symbol denoting a thematic role and x and y are terms, then V(x,y) is a basic condition;
- If x and y are terms, then $x{=}y$, $x{\neq}y$, $x{\sim}y$, $x{<}y$, $x{\leq}y$, $x{\prec}y$, and $x{\bowtie}y$ are basic conditions formed with comparison operators.

WordNet concepts are represented as word.`POS`.`SenseNum`, denoting a unique synset within WordNet. Thematic roles, including the VerbNet roles, always have two arguments and start with an uppercase character. Complex conditions introduce scopes in the meaning representation. They are defined using logical operators as follows:

24/3221: No one can resist. 00/2302: È tutto nuovo. 00/3008: Hij speelde piano en zij zong.




| | |
|---|---|
| **k0** | NOT **b2** |
| **b2** | REF **x1** |
| **b2** | person n.01 **x1** |
| **b2** | POS **b3** |
| **b3** | Agent **e1** **x1** |
| **b3** | REF **e1** |
| **b3** | resist v.02 **e1** |

| | |
|---|---|
| **k0** | IMP **b2** **b3** |
| **b2** | REF **x1** |
| **b2** | thing n.12 **x1** |
| **b3** | REF **s1** |
| **b3** | Theme **s1** **x1** |
| **b3** | new a.01 **s1** |
| **b3** | Time **s1** **t1** |
| **b4** | REF **t1** |
| **b4** | time n.08 **t1** |
| **b4** | EQU **t1** "now" |

| | | | |
|---|---|---|---|
| **k0** | DRS **k1** | **k0** | DRS **k2** |
| **b1** | REF **x1** | **b4** | REF **x3** |
| **b1** | male n.02 **x1** | **b4** | female n.02 **x3** |
| **k1** | REF **e1** | **k2** | REF **e2** |
| **k1** | play v.03 **e1** | **k2** | sing v.01 **e2** |
| **k1** | Agent **e1** **x1** | **k2** | Agent **e2** **x3** |
| **k1** | Theme **e1** **x2** | **b5** | REF **t2** |
| **k1** | REF **x2** | **b5** | time n.08 **t2** |
| **k1** | piano n.01 **x2** | **b5** | TPR **t2** "now" |
| **b3** | REF **t1** | **k2** | Time **e2** **t2** |
| **b3** | time n.08 **t1** | **k0** | CONTINUATION **k1** **k2** |
| **b3** | TPR **t1** "now" | **k1** | Time **e1** **t1** |

Figure 1: Examples of PMB documents with their scoped meaning representations and the corresponding clausal form. The first two structures are basic DRSs while the last one is a segmented DRS.

- If B is a DRS, then ¬B, ◇B, □B are complex conditions;
- If x is a variable, and B is a DRS, then x:B is a complex condition;
- If B and B' are DRSs, then B⇒B' and B∨B' are complex conditions.

Besides basic DRSs, we also have segmented DRSs, following Asher (1993) and Asher and Lascarides (2003). Hence, DRSs are formally defined as follows:

- If D is a (possibly empty) set of discourse referents, and C a (possibly empty) set of DRS-conditions, then <D,C> is a (basic) DRS;
- If B is a (basic) DRS, and B' a DRS, then B↓B' is a (segmented) DRS;
- If U is a set of labelled DRSs, and R a set of discourse relations, then <U,R> is a (segmented) DRS.

DRSs can be visualized in different ways. While the compact linear format saves space, the box notation increases readability. In this paper we use the latter notation. The examples of DRSs in the box notation are presented in Figure 1.

However, for evaluation and comparison purposes, we convert a DRS into a flat clausal form, i.e. a set of clauses. This is carried out by using the labels for DRSs as introduced in Venhuizen (2015) and Venhuizen et al. (2018), and breaking down the recursive structure of DRS by assigning them a label of the DRS in which they appear. Let t, t', and t" be meta-variables ranging over DRSs or terms. Let $\mathcal{C}$ be a set of WordNet concepts, $\mathcal{T}$ a set of the thematic roles, and $\mathcal{O}$ the set of DRS operators (REF, NOT, POS, NEC, EQU, NEQ, APX, LES, LEQ, TPR, TAB, IMP, DIS, PRP, DRS). The resulting clauses are then of the form t R t' or t R t' t" where R $\in \mathcal{C} \cup \mathcal{T} \cup \mathcal{O}$. The result of translating DRSs to sets of clauses is shown in Figure 1. In a clausal form, it is assumed that different variables are represented with different variable names and vice versa. Due to this, before translating a DRS to a clausal form, different discourse referents in the DRS must be represented with different variable names. This assumption significantly simplifies the matching process between clausal forms (Section 4) and makes it possible to recover the original box notation of a DRS from its clausal form.

### 2.2 Comparing DRSs to AMRs

Since DRSs in a clausal form come close to the triple notation of AMRs (Cai and Knight, 2013), and both aim to model meaning of natural language expressions, it is instructive to compare these two meaning representations. The main difference between AMRs and DRSs is that the latter ones have explicit scopes (boxes) and scopal operators such as negation. Due to the presence of scope in DRSs, their clauses are more complex than AMR triples. The length of DRS clauses varies from three to four, in contrast to the constant length of AMR triples. Additionally, DRS clauses contain two different types of variables, for scopes and discourse referents, whereas AMR triples have just one type.

Table 1: Distribution of clause types for 2,049 gold DRSs.

| Type | Description | Example | Total |
|---|---|---|---|
| REF | Discourse referent | `b3 REF x2` | 7,592 |
| NOT | Negation | `b1 NOT b2` | 204 |
| POS | Possibility (◊) | `b4 POS b5` | 55 |
| NEC | Necessity (□) | `b2 NEC b3` | 14 |
| IMP | Implication (⇒) | `b1 IMP b2 b3` | 104 |
| PRP | Proposition (:) | `b1 PRP x6` | 50 |
| REL | Discourse relation | `b1 CONTINUATION b2` | 71 |
| DRS | DRS as a condition | `b4 DRS b5` | 84 |
| Compare | Comparison operators | `x1 APX x2` | 2,100 |
| Concept | WordNet senses | `b2 hurt v.02 e3` | 7,545 |
| Role | Semantic roles | `b2 Agent e3 x4` | 7,516 |

Unlike AMRs, DRSs model tense. In general, the tense related information is encoded in a clausal form with three additional clauses, which express a WordNet concept, semantic role and a comparison operator. In order to give an intuition about the diversity of clauses in DRSs, Table 1 shows a distribution of various types of clauses in a corpus of DRSs (see Section 3). Since every logical operator carries a scope, their number represents a lower bound of the number of scopes in the meaning representations. In addition to logical operators, scopes are introduced by presupposition triggers like proper names or pronouns.

To make a meaningful comparison between AMRs and DRSs in terms of size, we compare the DRSs of 250,000 English sentences from the Parallel Meaning Bank (PMB; Abzianidze et al., 2017) to AMRs of the same sentences, produced by the state-of-the-art AMR parser from van Noord and Bos (2017). Statistics of the comparison are shown in Figure 2. On average, DRSs are about twice as large as AMRs, in terms of the number of clauses as well as the number of unique variables. This is obviously due to the explicit presence of scope in the meaning representation. However, for both meaning representations the number of clauses and variables increase linearly with sentence length.

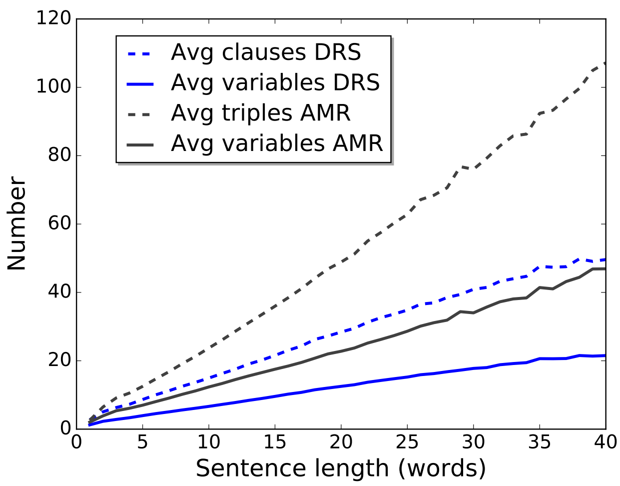



Figure 2: Comparison of the number of triples/clauses and variables between AMRs and DRSs for sentences of different length.

Table 2: Statistics of the first PMB release.

| | Documents | Sentences | Tokens |
|---|---|---|---|
| **English** | 2,049 | 2,057 | 11,664 |
| **German** | 641 | 642 | 3,430 |
| **Italian** | 387 | 387 | 1,944 |
| **Dutch** | 394 | 395 | 2,268 |

## 3 The Parallel Meaning Bank

The scoped meaning representations, integrating word senses, thematic roles, and the list of operators, form the final product of our semantically annotated corpus: the Parallel Meaning Bank. The PMB is a semantically annotated corpus of English texts aligned with translations in Dutch, German and Italian (Abzianidze et al., 2017). It uses the same framework as the Groningen Meaning Bank (Bos et al., 2017), but aims to abstract away from language-specific annotation models. There are five annotation layers present in the PMB: segmentation of words, multi-word expressions and sentences (Evang et al., 2013), semantic tagging (Bjerva et al., 2016; Abzianidze and Bos, 2017), syntactic analysis based on CCG (Lewis and Steedman, 2014), word senses based on WordNet (Fellbaum, 1998), and thematic role labelling (Bos et al., 2012). The semantic analysis for English is projected on the other languages, to save manual annotation efforts (Evang, 2016; Evang and Bos, 2016). All the information provided by these layers is combined into a single meaning representation using the semantic parser Boxer (Bos, 2015), in the form of Discourse Representation Structures. Note that the goal is to produce annotations that capture the most probable interpretation of a sentence; no ambiguities or under-specification techniques are employed.

At each step in this pipeline, a single component produces the automatic annotation for all four languages, using language-specific models. Human annotators can correct machine output by adding 'Bits of Wisdom' (Basile et al., 2012). These corrections serve as data for training better models, and create a gold standard annotated subset of the data. Annotation quality is defined per layer and language, at three levels: bronze (fully automatic), silver (automatic with some manual corrections), and gold (fully manually checked and corrected). If all layers are marked as gold, it follows that the resulting DRS can be considered gold standard, too.

The first public release[1] of the PMB contains gold standard scoped meaning representations for over 3,000 sentences in total (see Table 2). The release includes mainly relatively short sentences involving several semantic scope phenomena. A detailed distribution of clause types in the dataset is given in Table 1. A larger amount of texts and more complex linguistic phenomena will be included in future releases.

In addition to the released data, the PMB documents are publicly accessible through a web interface, called the PMB explorer.[2] In the explorer, visitors can view natu-

[1] `http://pmb.let.rug.nl/data.php`
[2] `http://pmb.let.rug.nl/explorer`

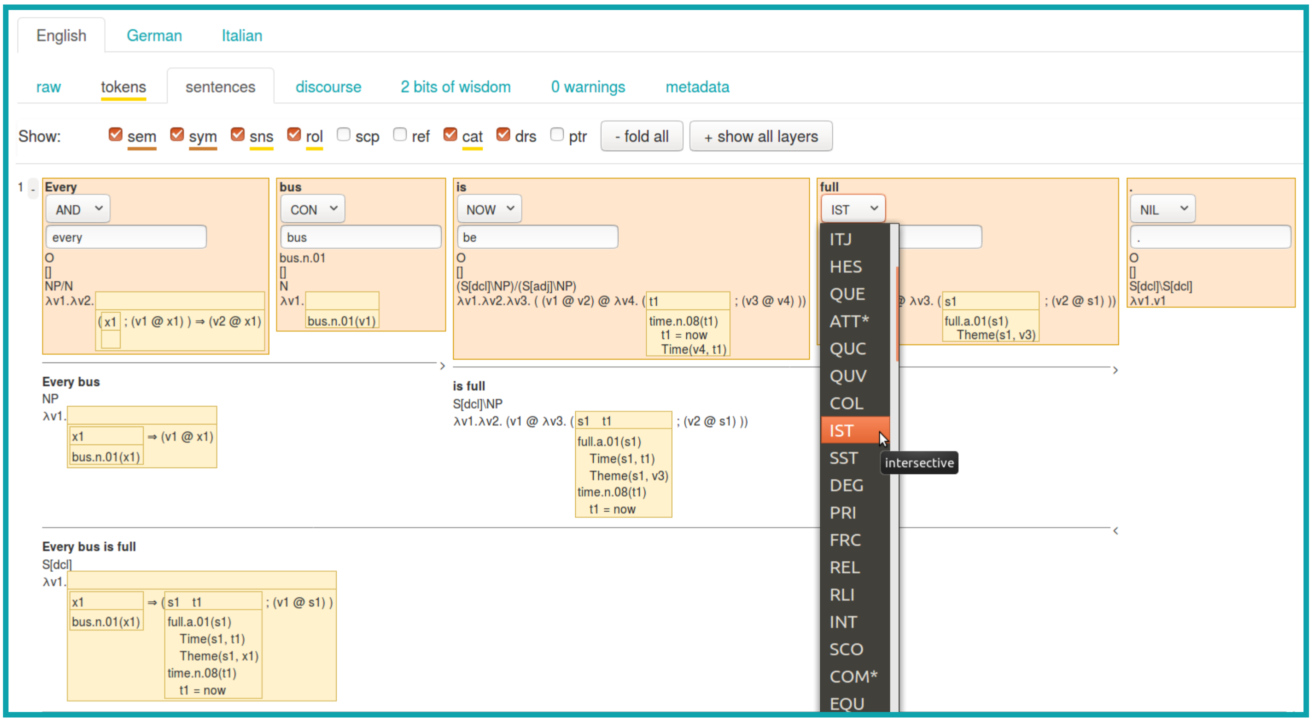



Figure 3: The edit mode of the PMB explorer: semantic tag (`sem`) and symbol (`sym`) layers of the document are bronze and therefore editable, while the word sense (`sns`), semantic role (`rol`) and CCG category (`cat`) layers are gold and uneditable.

ral language texts with several layers of annotations and compositionally derived meaning representations, and, after registration, edit the annotations. It is also possible to use a word or a phrase search to find certain words or constructions with their semantic analyses. Figure 3 shows the PMB explorer with the semantic analysis of a sentence in the edit mode.

## 4 Matching Scoped Representations

### 4.1 Evaluation by Matching

In the context of the Parallel Meaning Bank there are two main reasons to verify whether two scoped meaning representations capture the same meaning or not: (1) to be able to evaluate semantic parsers that produce scoped meaning representations by comparing gold-standard DRSs to system output; and (2) to check whether translations are meaning-preserving; a discrepancy in meaning between source and target could indicate a mistranslation.

The ideal way to compare two meaning representations would be one based on inference. This can be implemented by translating DRSs to first-order formulas and using an off-the-shelf theorem prover to find out whether the two meanings are logically equivalent (Blackburn and Bos, 2005). This method can compare meaning representation that have different syntactic structures but still are equivalent in meaning. The disadvantage of this approach is that it yields just a binary answer: if a proof is found the meanings are the same, else they are not.

An alternative way of comparing meaning representations is comparing the corresponding clausal forms by computing precision and recall over matched clauses (Allen et al., 2008). The advantage of this approach is that it returns a score between 0 and 1, preferring meaning representations that better approximate the gold standard over those that are completely different. Since the variables of different clausal forms are independent from each other, the comparison of two clausal forms boils down to finding a (partial) one-to-one variable mapping that maximizes intersection of the clausal forms. For example, the maximal matching for the clausal forms in Figure 4 is achieved by the following partial mapping from the variables of the left form into the variables of the right one: $\{\texttt{k0}\mapsto\texttt{b0}, \texttt{e1}\mapsto\texttt{v1}\}$.

For AMRs, finding a maximal matching is done using a hill-climbing algorithm called SMATCH (Cai and Knight, 2013). This algorithm is based on a simple principle: it checks if a single change in the current mapping results in a better matching mapping. If this is the case, it continues with the new mapping. Otherwise, the algorithm stops and has arrived at the final mapping. This means that it can easily get stuck in local optima. To avoid this, SMATCH does a predefined number of restarts of this process, where each restart starts with a new and random initial mapping. The first restart always uses a 'smart' initial mapping, based on matching concepts.

Our evaluation system, called COUNTER[3], is a modified version of SMATCH. Even though clausal forms do not form a graph and clauses consist of either three or four components, the principle behind the variable matching is the same. The actual implementation differs, mainly because SMATCH was not designed to handle clauses with three variables, e.g. $\langle\texttt{k0 Agent e1 x1}\rangle$.

In contrast to SMATCH, COUNTER takes a set of clauses directly as input. COUNTER also uses two smart initial map-

[3] `http://github.com/RikVN/DRS_parsing/`

01/3445: `He smiled.`

| $x_1$ $e_1$ $t_1$ |
|---|
| male.n.02($x_1$)<br>smile.v.01($e_1$)<br>Time($e_1,t_1$)<br>Agent($e_1,x_1$)<br>time.n.08($t_1$)<br>$t_1 \prec$ now |

SPAR DRS

~~b1 REF x1~~
b1 male n.02 x1
~~b3 REF t1~~
b3 TPR t1 "now"
b3 time n.08 t1
k0 Agent e1 x1
~~k0 REF e1~~
k0 Time e1 t1
k0 smile v.01 e1

00/3514: `She fled Australia.`

| $x_1$ $x_2$ $v_1$ $t_1$ |
|---|
| female.n.02($x_1$)<br>flee.v.01($v_1$)<br>Time($v_1,t_1$)<br>Source($v_1,x_2$)<br>Theme($v_1,x_1$)<br>time.n.08($t_1$)<br>$t_1 \prec$ now<br>country.n.02($x_2$)<br>Name($x_2$, australia) |

~~b1 REF x1~~
b1 female n.02 x1
~~b3 REF t1~~
b3 TPR t1 "now"
b3 time n.08 t1
b0 Theme v1 x1
b0 Source v1 x2
~~b0 REF v1~~
b0 Time v1 t1
b0 flee v.01 v1
~~b2 REF x2~~
b2 Name x2 "australia"
b2 country n.02 x2

Figure 4: The SPAR DRS (Section 5.1) matches the DRS of 00/3514 PMB document with an F-score of 54.5%. If redundant REF-clauses are ignored, the F-score drops to 40%. These results are achieved with the help of the mapping {`k0`↦`b0`, `e1`↦`v1`}.

pings, based on either role-clauses, like ⟨`k0 Agent e1 x1`⟩, or concept-clauses, like ⟨`k0 smile v.01 e1`⟩.

Also specific to this method is the treatment of REF-clauses in the matching process. Before matching two DRSs, redundant REF-clauses are removed. A REF-clause ⟨`b1 REF x1`⟩ is redundant if its discourse referent `x1` occurs in some basic condition of the same DRS `b1`. Figure 4 shows some examples of redundant REF-clauses. Not removing these redundant clauses would lead to inflated matching scores since for each matched variable the corresponding REF-clause will also match. Comparison of the clausal forms in Figure 4 demonstrates this fact. Note that not all REF-clauses are redundant: if a discourse referent is declared outside the scope of negation or an other scope operator, the REF-clause is kept. This is very infrequent in our data, since only a single REF-clause was preserved in 2,049 examples.

### 4.2 Evaluating Matching

As we showed in Figure 2, DRSs are about twice as large as AMRs. This increase in size might be problematic, since it increases the average runtime for comparing DRSs. Moreover, if there are more variables, more restarts might be needed to ensure a reliable score, again increasing runtime.

Therefore, our goal is that COUNTER gets close to optimal performance in reasonable time. Since we want to be sure that this also holds for longer sentences, we use a balanced data set. We take 1,000 DRSs produced by the semantic parser Boxer for each sentence length from 2 to 20 (punctuation excluded), resulting in a set of 19,000 DRSs.

`She removed the dishes from the table.`

```
(r / remove-01
   :ARG0 (s / she)
   :ARG1 (d / dish)
   :ARG2 (t / table))
```

```
b0 REF x1
b0 remove v.01 x1
b4 REF x5
b4 TPR x5 "now"
b4 time n.08 x5
b0 Time x1 x5
b0 Agent x1 x2
b1 REF x2
b1 female n.02 x2
b0 Patient x1 x3
b2 REF x3
b2 dish n.01 x3
b0 Theme x1 x4
b3 REF x4
b3 table n.01 x4
```

Figure 5: A clausal form obtained from an automatically generated AMR of the document 14/0849.

To test COUNTER in a realistic setting, we cannot compare the DRSs to themselves or to a DRS of the translation, since those are too similar. Therefore, the 19,000 English sentences of the DRS are parsed by an existing AMR parser (van Noord and Bos, 2017) and subsequently converted into a DRS by a rule-based system, AMR2DRS, as motivated by Bos (2016). An example of translating an AMR to a clausal form of a DRS is shown in Figure 5. We convert AMR relations to DRS roles by employing a manually created translation dictionary, including rules for semantic roles (e.g. `:ARG0` ↦ `Agent` and `:ARG1` ↦ `Patient`) and pronouns (e.g. `she` ↦ `female.n.02`). Since AMRs do not contain tense information, past tense clauses[4] are produced for the first verb in the AMR (see four tense related clauses in Figure 5). Also, since AMRs do not use WordNet synsets, all concepts get a default first sense, except for concepts that are added by concept-specific rules, such as `female.n.02` and `time.n.08`.

We compare the sets of DRSs using different numbers of restarts to find the best trade-off between speed and accuracy. The results are shown in Table 3. The optimal scores are obtained using a Prolog script that performs an exhaustive search for the optimal mapping. As expected, increasing the number of restarts benefits performance. Cai and Knight (2013) consider four restarts the optimal trade-off between accuracy and speed, showing no improvement in F-score when using more than ten restarts.[5] Contrary to SMATCH, performance for COUNTER still increases with more than 4 restarts. In our case, it is a bit harder to select an optimal number of restarts, since this number depends on the length of the sentence, as shown in Figure 6. We see that for long sentences, 5 and 10 restarts are not sufficient to get close to the optimal, while for short sentences 5 restarts might be considered enough. In general, the best trade-off between speed and accuracy is approximately 20 restarts.

[4] Past tense was chosen because it is the most frequent tense in the data set.

[5] However, we found that, in practice, SMATCH still improves when using more restarts. Parsing the development set of the AMR dataset LDC2016E25 with the baseline parser of van Noord and Bos (2017) yields an F-score of 55.0 for 10 restarts, but 55.4 for 100 restarts.

Table 3: Results of comparing 19,000 Boxer-produced DRSs to DRSs produced by AMR2DRS, for different number of restarts. For three or more restarts, we always use the smart role and concept mapping.

| Restarts | P% | R% | F1% | Time (h:m:s) |
|---|---|---|---|---|
| (random) 1 | 27.20 | 22.71 | 24.75 | 4:19 |
| (smart concepts) 1 | 27.45 | 22.92 | 24.98 | 4:35 |
| (smart roles) 1 | 27.27 | 22.76 | 24.81 | 4:37 |
| 5 | 30.25 | 25.25 | 27.53 | 19:33 |
| 10 | 30.65 | 25.59 | 27.89 | 37:08 |
| 20 | 30.84 | 25.75 | 28.07 | 1:10:13 |
| 30 | 30.90 | 25.80 | 28.12 | 1:41:43 |
| 50 | 30.94 | 25.83 | 28.16 | 2:41:38 |
| 75 | 30.96 | 25.85 | 28.17 | 3:53:01 |
| 100 | 30.97 | 25.85 | 28.18 | 5:01:25 |
| Optimal | 30.98 | 25.86 | 28.19 | |

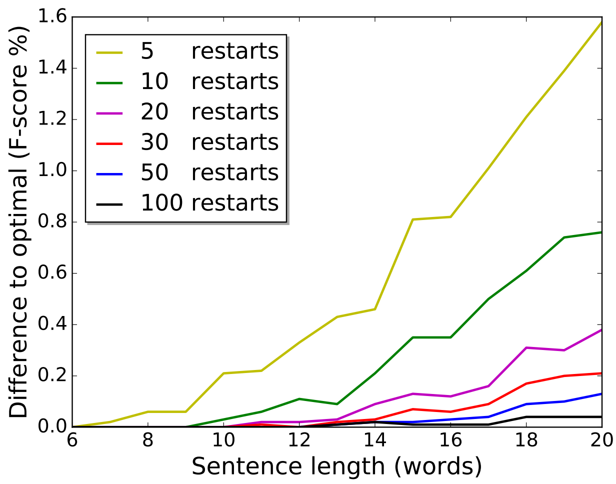



Figure 6: Comparison of the differences to the optimal F-score per sentence length for different number of restarts.

## 5 COUNTER in Action

### 5.1 Semantic Parsing

The first purpose of COUNTER is to evaluate semantic parsers for DRSs. Since this is a new task, there are no existing systems that are able to do this. Therefore, we show the results of three baseline systems PMB PIPELINE, SPAR, and AMR2DRS (Subsection 4.2).[6]

The PMB PIPELINE produces a DRS via the pipeline of the tools used for automatic annotation of the PMB.[7] This means that it has no access to manual corrections, and hence it uses the most frequent word senses and default VerbNet roles. SPAR is a trivial semantic 'parser' which always outputs the DRS that is most similar to all other DRSs in the most recent PMB release (the left-hand DRS in Figure 4).

The results of the three baseline parsers are shown in Table 4. The surprisingly high score of SPAR is explained by the fact that the first PMB release mainly contains relatively short sentences with little structural diversity. The average number of clauses per clausal form (excluding redundant REF-clauses) is 8.7, where a substantial share (approximately 3) comes from tense related clauses. Due to this fact, guessing temporal clauses for short sentences has a big impact on F-score. This is illustrated by the comparison of the clausal forms in Figure 4, where matching only temporal clauses results in an F-score of 40%.

[6]SPAR and AMR2DRS are available at: `https://github.com/RikVN/DRS_parsing/`

[7]`http://pmb.let.rug.nl/software.php`

Table 4: Comparison of three baseline DRS parsers to the gold-standard data set.

| | Precision% | Recall% | F-score% |
|---|---|---|---|
| SPAR | 53.1 | 36.6 | 43.3 |
| AMR2DRS | 46.5 | 48.2 | 47.3 |
| PMB PIPELINE | 53.0 | 54.8 | 53.9 |

AMR2DRS outperforms SPAR by a considerable margin, but is still far from optimal. This is also the case for PMB PIPELINE, which shows that, within the PMB, manual annotation is still required to obtain gold standard meaning representations.

### 5.2 Comparing Translations

The second purpose of COUNTER is checking whether translations are meaning-preserving. As a pilot study, we compare the gold standard meaning representations of German, Italian and Dutch translations in the release to their English counterparts. The results are shown in Table 5. The high F-scores indicate that the meaning representations are often syntactically very similar, if not identical. However, there is a considerable subset of meaning representations which are different from the English ones, indicating that there is at least a slight discrepancy in meaning for those translations.

Table 5: Comparing meaning representations of English texts to those of German, Italian and Dutch translations.

| | F-score% | Docs | F<1.0 | % total |
|---|---|---|---|---|
| **German** | 98.4 | 579 | 61 | 10.5 |
| **Italian** | 97.6 | 341 | 46 | 13.5 |
| **Dutch** | 98.3 | 355 | 37 | 10.4 |

Manual analysis of these discrepancies showed that there are several different causes for a discrepancy to arise. In most of the cases (38%), a human annotation error was made. In 34% of cases, a definite description was used in one language but not in the other. Examples are 'has long hair' with the Italian translation 'ha i capelli lunghi', and 'escape from prison' with the Dutch translation 'vluchtte uit de gevangenis'. In 15% of cases proper names were translated (e.g. 'United States' and 'Stati Uniti'). This is not accounted for, since we do not currently make use of grounding proper names to a unique identifier, for instance by wikification (Cucerzan, 2007), or by using a language-independent transliteration of names. In 13% of cases the translation was either non-literal or incorrect. Examples are 'Tom lacks experience' with the Dutch translation 'Tom

She removed the dishes from the table.

| $x_1$ $x_2$ $e_1$ $x_3$ $t_1$ |
|---|
| female.n.02($x_1$) |
| remove.v.01($e_1$) |
| Time($e_1, t_1$) |
| Source($e_1, x_3$) |
| Theme($e_1, x_2$) |
| Agent($e_1, x_1$) |
| time.n.08($t_1$) |
| $t_1 \prec$ now |
| dish.n.01($x_2$) |
| table.n.03($x_3$) |

~~b1 REF x1~~
b1 female n.02 x1
~~b5 REF t1~~
b5 TPR t1 "now"
b5 time n.08 t1
k0 Agent e1 x1
~~k0 REF e1~~
k0 Theme e1 x2
k0 Time e1 t1
k0 remove v.01 e1
~~b2 REF x2~~
b2 dish n.01 x2
k0 Source e1 x3
~~b4 REF x3~~
b4 table n.03 x3

Ze ruimde de tafel af.

| $x_1$ $x_2$ $e_1$ $t_1$ |
|---|
| female.n.02($x_1$) |
| unclutter.v.01($e_1$) |
| Time($e_1, t_1$) |
| Source($e_1, x_2$) |
| Agent($e_1, x_1$) |
| time.n.08($t_1$) |
| $t_1 \prec$ now |
| table.n.03($x_2$) |

~~b1 REF x1~~
b1 female n.02 x1
~~b4 REF t1~~
b4 TPR t1 "now"
b4 time n.08 t1
k0 Agent e1 x1
~~k0 REF e1~~
k0 Source e1 x2
k0 Time e1 t1
k0 unclutter v.01 e1
~~b2 REF x2~~
b2 table n.03 x2

Figure 7: English and Dutch non-literal translations of the document 14/0849. Their clausal forms match each other (excl. redundant REF-clauses) with an F-score of 77.8%. This matching is achieved by the mapping of variables {b5↦b4, b4↦b2}.

heeft geen ervaring' (lit. 'Tom has no experience'), 'can't use chopsticks' with the German 'kann nicht mit Stäbchen essen' (lit. 'cannot eat with sticks'), and 'remove the dishes from the table' with the Dutch translation 'ruimde de tafel af' (lit. 'uncluttered the table').

The mapping of clausal forms involving non-literal translations is illustrated in Figure 7. This preliminary analysis shows that this comparison of meaning representations provides an an additional method for detecting mistakes in annotation. It also showed that there are cases where our semantic analysis needs to be revised and improved.

## 6 Conclusions and Future Work

Large semantically annotated corpora are rare. Within the Parallel Meaning Bank project, we are creating a large, open-domain corpus annotated with formal meaning representations. We take advantage of parallel corpora, enabling the production of meaning representations for several languages at the same time. Currently, the these are languages similar to English, two Germanic languages (Dutch and German) and one Romance language (Italian). Ideally, future work would include more non-Germanic languages.

The DRSs that we present are meaning representations with substantial expressive power. They deal with negation, universal quantification, modals, tense, and presupposition. As a consequence, semantic parsing for DRSs is a challenging task. Compared to Abstract Meaning Representations, the number of clauses and variables in a DRS is about two times larger on average. Moreover, compared to AMRs, DRSs rarely contain clauses with single variables. All non-logical symbols used in DRSs are grounded in WordNet and VerbNet (with a few extensions). This makes evaluation using matching computationally challenging, in particular for long sentences, but our matching system COUNTER achieves a reasonable trade-off between speed and accuracy.

Several extensions to the annotation scheme are possible. Currently, the DRSs for the non-English languages contain references to synsets of the English WordNet. Conceptually, there is nothing wrong with this (as synsets can be viewed as identifiers for concepts that are language-independent), but for practical reasons it makes more sense to provide links to synsets of the original language (Hamp and Feldweg, 1997; Postma et al., 2016; Roventini et al., 2000; Pianta et al., 2002). In addition, we consider implementing semantic grounding such as wikification in the Parallel Meaning Bank.

As for other future work, we plan to include a more fine-grained matching regarding WordNet synsets, since the current evaluation of concepts is purely string-based, with only identical strings resulting in a matching clause. For many synsets, however, it is possible to refer to them with more than one word.POS.SenseNum triple, and this should be accounted for (e.g. fox.n.02 and dodger.n.01 both refer to the same synset). In a similar vein, we plan to experiment with including WordNet concept similarity techniques in COUNTER to compute semantic distances between synsets, in case they do not fully match.

Finally, we would like to stimulate research on semantic parsing with scoped meaning representations. Not only are we planning to extend the coverage of phenomena and the number of texts with gold-standard meaning representations for the four languages, we also aim to organize a shared task on DRS parsing for English, German, Dutch and Italian in the near future.

## 7 Bibliographical References


Abzianidze, L. and Bos, J. (2017). Towards universal semantic tagging. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017) – Short Papers*, Montpellier, France, September. Association for Computational Linguistics.

Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., van Noord, R., Ludmann, P., Nguyen, D.-D., and Bos, J. (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: Volume 2, Short Papers*, pages 242–247, Valencia, Spain, April. Association for Computational Linguistics.

Allen, J. F., Swift, M., and de Beaumont, W. (2008). Deep Semantic Analysis of Text. In Johan Bos et al., editors, *Semantics in Text Processing. STEP 2008 Conference*

*Proceedings*, volume 1 of *Research in Computational Semantics*, pages 343–354. College Publications.

Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Studies in natural language processing. Cambridge University Press.

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.

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

Basile, V., Bos, J., Evang, K., and Venhuizen, N. (2012). 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 2012)*, pages 92–96, Avignon, France.

Bjerva, J., Plank, B., and Bos, J. (2016). Semantic tagging with deep residual networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3531–3541, Osaka, Japan.

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

Bonial, C., Corvey, W. J., Palmer, M., Petukhova, V., and Bunt, H. (2011). A hierarchical unification of LIRICS and VerbNet semantic roles. In *Proceedings of the 5th IEEE International Conference on Semantic Computing (ICSC 2011)*, pages 483–489.

Bos, J., Evang, K., and Nissim, M. (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.

Bos, J., Basile, V., Evang, K., Venhuizen, N., and Bjerva, J. (2017). The Groningen Meaning Bank. In Nancy Ide et al., editors, *Handbook of Linguistic Annotation*, volume 2, pages 463–496. Springer.

Bos, J. (2015). Open-domain semantic parsing with Boxer. In Beáta Megyesi, editor, *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 301–304.

Bos, J. (2016). Expressive power of Abstract Meaning Representations. *Computational Linguistics*, 42(3):527–535.

Bos, J. (2017). Indexicals and compositionality: Inside-out or outside-in? In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017) – Short Papers*, Montpellier, France, September. Association for Computational Linguistics.

Cai, S. and Knight, K. (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, August. Association for Computational Linguistics.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

Evang, K. and Bos, J. (2016). Cross-lingual learning of an open-domain semantic parser. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 579–588, Osaka, Japan.

Evang, K., Basile, V., Chrupała, G., and Bos, J. (2013). Elephant: Sequence labeling for word and sentence segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1422–1426, Seattle, Washington, USA.

Evang, K. (2016). *Cross-lingual Semantic Parsing with Categorial Grammars*. Ph.D. thesis, University of Groningen.

Christiane Fellbaum, editor. (1998). *WordNet. An Electronic Lexical Database*. The MIT Press, Cambridge, Ma., USA.

Hamp, B. and Feldweg, H. (1997). GermaNet - a lexical-semantic net for German. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT*. Kluwer, Dordrecht.

Lewis, M. and Steedman, M. (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.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302.

Postma, M., van Miltenburg, E., Segers, R., Schoen, A., and Vossen, P. (2016). Open Dutch WordNet. In *Proceedings of the Eight Global Wordnet Conference*, Bucharest, Romania.

Roventini, A., Alonge, A., Calzolari, N., Magnini, B., and Bertagna, F. (2000). ItalWordNet: a large semantic database for Italian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2000)*, pages 783–790.

van Noord, R. and Bos, J. (2017). Neural semantic parsing by character-based translation: Experiments with Abstract Meaning Representations. *Computational Linguistics in the Netherlands Journal*, 7:93–108.

Venhuizen, N. J., Bos, J., Hendriks, P., and Brouwer, H. (2018). Discourse semantics with information structure. *Journal of Semantics*.

Venhuizen, N. J. (2015). *Projection in Discourse: A data-driven formal semantic analysis*. Ph.D. thesis, University of Groningen.