

A new well-formedness criterion for semantics debugging

Dan Flickinger¹, Alexander Koller², Stefan Thater²

¹ CSLI, Stanford University, Stanford

² Universität des Saarlandes, Saarbrücken

Proceedings of the HPSG 05 Conference

Department of Informatics, University of Lisbon

Stefan Müller (Editor)

2005

CSLI Publications

<http://csli-publications.stanford.edu/>

Abstract

We present a novel well-formedness condition for underspecified semantic representations which requires that every correct MRS representation must be a *net*. We argue that (almost) all correct MRS representations are indeed nets, and apply this condition to identify a set of eleven rules in the English Resource Grammar (ERG) with bugs in their semantics component. Thus we demonstrate that the net test is useful in grammar debugging.

1 Introduction

A very exciting recent development in (computational) linguistics is that large-scale grammars which derive semantic representations for their input sentences are becoming available. For instance, the English Resource Grammar (Copestake and Flickinger, 2000) is a large-scale HPSG grammar for English which computes underspecified semantic representations in the MRS formalism (Copestake et al., 2004). It is standard to use underspecification to deal with scope ambiguities; apart from MRS, there is a number of other underspecification formalisms, such as dominance constraints (Egg et al., 2001) and Hole Semantics (Bos, 1996).

However, the increased power of the new grammars comes with a new challenge for grammar engineering: How can we be sure that all semantic outputs the grammar computes (through any combination of semantic construction rules) are correct, and how can we find and fix bugs? This problem of *semantics debugging* is an important factor in the 90% of grammar development time that is spent on the syntax-semantics interface (Copestake et al., 2001).

Grammar development systems such as the LKB implement some semantic sanity checks, which are practically useful, but rather shallow, and therefore limited in their power. On the theoretical side, there are attempts to formalise “best practices” of grammar development in a *semantic algebra* (Copestake et al., 2001), but this is quite a far-reaching project that is not yet fully implemented.

One potential alternative method for semantics debugging comes from Fuchss et al.’s recent work on *nets* (Fuchss et al., 2004). They claim that every underspecified description (written in MRS or as a dominance constraint) that is actually used in practice is a *net*, i.e. it belongs to a restricted class of descriptions with certain useful structural properties, and they substantiate their claim through an empirical evaluation on a treebank. We report further evidence for this “Net Hypothesis” here by investigating the only three non-nets in the ERG’s Semantic Test Suite in some more detail. If the Net Hypothesis is true, we can recognise a grammar rule (or combination of rules) as problematic if it produces non-nets.

In this paper, we show that such a use of nets is indeed possible. We collect all MRSs that the ERG derives for all sentences in the Rondane treebank (distributed with the ERG) and the Verbmobil sections of the Redwoods treebank (Oepen et al., 2002). Then we look for the grammar rules that are responsible for deriving the non-nets, and identify a group of eleven rules which only produce non-nets for any sentence in whose analysis they are involved. By manually inspecting these

eleven rules, we determine that they indeed all have faulty semantics components. We have manually corrected some of these rules, and the corrections have been incorporated into newer versions of the ERG.

Plan of the paper. We will first give a brief definition of MRS in Section 2 and of MRS nets in Section 3. Then we will state the Net Hypothesis and report evidence for it in Section 4. The core of the paper is Section 5, in which we show how we can identify semantically buggy grammar rules by looking for non-nets in corpus data. Section 6 concludes the paper and points to future work.

2 Minimal Recursion Semantics

We start with a an informal overview of Minimal Recursion Semantics (MRS) – for details see (Copestake et al., 2004) and (Fuchss et al., 2004). MRS is the standard scope underspecification formalism used in current HPSG grammars, such as the English Resource Grammar (Copestake and Flickinger, 2000) or grammars derived from the Grammar Matrix (Bender et al., 2002). Its purpose is to separate the problem of resolving scope ambiguities from semantics construction.

Abstract Syntax. An *MRS structure*, or simply *MRS* for short, consists of a set of *elementary predications (EPs)* and *handle constraints*. Elementary predications can be thought of as “labeled” first order formulas with “holes.” The idea is that an MRS describes a set of first order formulas that one can obtain by “plugging” labels into holes, while handle constraints restrict possible pluggings. Consider for instance the following MRS for the sentence “each section is also suitable as a single day tour” from the Rondane treebank:

$$\{l_0 : \text{proposition}(h_1), l_2 : \text{udef}(x, h_3, h_4), l_5 : \text{a}(y, h_6, h_7), l_8 : \text{each}(z, h_9, h_{10}), \\ l_{11} : \text{single}(x), l_{11} : \text{day}(x), l_{12} : \text{tour}(y), l_{12} : \text{compound}(x, y), \\ l_{13} : \text{section}(z), l_{14} : \text{suitable}(z), l_{14} : \text{also}, l_{14} : \text{as}(y), \\ h_3 =_q l_{11}, h_6 =_q l_{12}, h_9 =_q l_{13}, h_{10} =_q l_{14}\}$$

Terms of the form $l : P(\dots)$ are elementary predications. l is the label of the EP, terms h on the right hand side of ‘:’ are *argument handles*, and terms x, y, \dots are ordinary first order variables. Terms of the form $h =_q l$ are handle constraints, also called *qeq-constraints*, which specify, approximately, that h must outscope l in all scope-resolved MRS structures (see below). Note that each label can label more than one EP (e.g. l_{12} in the example). This is called an *EP-conjunction* and is interpreted as a conjunction of the formulas labelled by l_{12} . Note also that first-order variables like z are bound in quantifier EPs (here, the one labelled by l_8) and used as bound variables in other EPs (such as the one labelled by l_{13}).

We usually represent MRS structures as graphs (see Fuchss et al., 2004). For instance, the MRS above can be represented by the graph in Fig. 1. The nodes of the graph are the labels and argument handles of the MRS, and the solid edges correspond to EPs. EP conjunctions are represented by explicit conjunction at the graph

EP conjunctions than the original MRS structure. In such a case, we call the scope-resolved MRS structure a *merging configuration*.

3 MRS-Nets

We now introduce (*MRS-*) *nets*, which are MRS structures that satisfy certain additional constraints.

We say that an MRS structure is an (MRS-) net if and only if every fragment in its graph satisfies the following two conditions:

1. There is exactly one node without outgoing dominance edges. All other nodes in the fragment have at least one outgoing dominance edge.
2. If a node X has two (or more) outgoing dominance edges, say, to Y and Z , then Y and Z are connected by a *hypernormal path* (see below) that does not visit the node X itself.

A *hypernormal path* in a graph is an undirected path that does not use two dominance edges that start from the same node. For instance, the following two paths are hypernormal:



By contrast, the following path is not hypernormal:



The MRS graph shown in Fig. 1 is an example of a net. The quantifier fragments all have a single node (the “scope” of the quantifier) without outgoing dominance edges, while all other nodes have exactly one outgoing dominance edge, so they satisfy the first net-conditions; the second net condition is trivially satisfied. This is also the case for the nuclear fragments l_{11}, l_{12}, \dots , that have no outgoing dominance edges at all. The only fragment with nodes that have more than one outgoing dominance edges is the top fragment. Its dominance children are the three quantifier fragments, and there is a hypernormal path between each pair of these fragments – for instance, l_2, l_{12}, h_6, l_5 and l_5, l_{14}, l_8 . Hence, all fragments satisfy the two net conditions.

On the other hand, Fig. 6 shows two MRS structures which are not nets because the top fragments violate the second net condition. For example, in the first graph the top fragment has dominance edges to the fragments for “a bit” and “two young Norwegians”. But the only (undirected) path that connects these two fragments goes through the top fragment itself, and this path is not hypernormal. This graph also contains a quantifier fragment (“a bit”) which has two nodes without outgoing dominance edges and thus violates the first net condition.

The definition above is a generalisation of the original definition of nets that we gave in earlier papers (Niehren and Thater, 2003). We use it because the earlier definition involves some rather arbitrary restrictions about the allowable fragments – for example, it excludes fragments whose root has two or more outgoing dominance edges. However, all statements about nets in (Niehren and Thater, 2003) remain true for the new definition, and the proofs carry over almost verbatim. In particular, the key theorem which motivated the definition of nets remains true:

Theorem 1 (Niehren and Thater (2003)). *If (the graph of) an MRS is a net, then the MRS can be translated into a normal dominance constraint such that the configurations of the MRS bijectively correspond to the solved forms of the corresponding dominance constraint.*

This means that nets can be solved efficiently using the solvers for normal dominance constraints (e.g., Bodirsky et al., 2004). But nets have useful formal properties even from a pure MRS perspective. For example, it can be shown that MRS nets never have merging configurations. This means that EP conjunctions can generally be resolved in a preprocessing step, and need never be dynamically introduced by the solver.

4 The Net Hypothesis

Beyond these formal properties, one intriguing aspect of nets is that it is extremely hard to find useful underspecified descriptions that are not nets. This made Fuchss et al. (2004) propose the following “net hypothesis”:

Net Hypothesis. *All underspecified descriptions (e.g., MRS structures) that are used in practice for scope underspecification are nets.*

This hypothesis looks surprising at first glance. The intuition is that the second net condition, in particular, reflects the fact that quantifiers in underspecified representations are derived from noun phrases that are arguments of predicates like verbs or prepositions. In the underspecified representation, the arguments of such a predicate are variables which are bound by the quantifiers, and because quantifiers must outscope the variables they bind, this creates a hypernormal path between the two quantifiers. For example, the two quantifiers for “we” and “two young Norwegians” in Fig. 6 bind the arguments of the verb “meet”. If the variable x , which is bound by the quantifier “a bit”, was used anywhere else in the MRS structure (as it should be), this use would create a hypernormal connection between this quantifier and the rest of the graph; and so on.

4.1 Previous Evidence

One approach to determining whether the Net Hypothesis is true is to look at large corpora of MRS structures and checking whether they are nets or not. Fuchss et al.

(2004) presented a first evaluation along these lines. They considered the sentences in the Redwoods treebank (Oepen et al., 2002) and generated the MRS structures for all syntactic analyses of these sentences according to the English Resource Grammar (ERG; Copestake and Flickinger, 2000). It turned out that about 83 % of the well-formed MRS structures obtained in this way were in fact nets, while about 17 % aren't.

In addition, they evaluated a number of these non-nets manually and found that they seemed to be systematically incomplete: The MRS graphs were missing some dominance or binding edges, with the consequence that they permitted semantic readings that the original sentences didn't have. This impression was further substantiated by the fact that the average number of configurations for the non-nets was about seven times higher than for the nets.

4.2 Experiments with the Semantic Test Suite

There are two possible explanations for the fact that 17 % of Fuchss et al.'s MRS structures were non-nets. One is that the Net Hypothesis is wrong, and a substantial number of these non-nets are legitimate underspecified representations. The other explanation is that the Net Hypothesis is in fact true, and the non-nets result from errors in the syntax-semantics interface of the grammar, which a system of the ERG's complexity can be expected to have.

In order to shed light on this question, we performed an experiment with the MRS structures in the *Semantic Test Suite* (STS), which is distributed with the ERG grammar. This is a collection of 107 hand-constructed sentences with syntactic and semantic annotation which is used as a test suite for debugging the MRS solver in the LKB system (Copestake and Flickinger, 2000). We expected that because the MRSs in this artificial corpus are routinely examined by hand and corrected, they should tend to be less sensitive to possible errors in the grammar than a corpus of real-world text such as Redwoods. If the Net Hypothesis is true, all MRSs in the STS should be nets.

We evaluated this for the October 2004 version of the STS. It turned out that of the 107 MRS structures, only three were not nets. These resulted from the following three sentences:

- (1) The dog barked, didn't it? (sentence 77; 3 scopings)
- (2) It took Abrams ten minutes to arrive. (83; 4 scopings)
- (3) How happy was Abrams? (102; 3 scopings)

Upon closer inspection, we claim that all three MRS structures (shown in Fig. 3) are still incomplete and should really be nets. This is most obvious for (2), whose MRS graph is shown on the bottom in Fig. 3. "Ten minutes" is an argument of "take", so it is a clear error that the predicate "take" is only applied to y and not to x (the variable bound by the quantifier for "ten minutes") in the MRS.

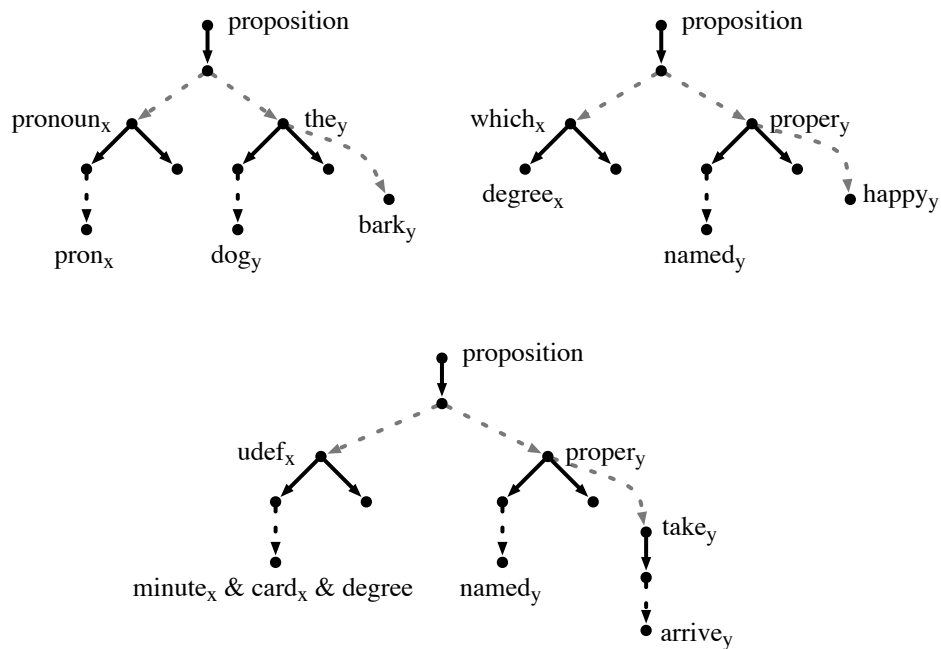


Figure 3: The three non-nets from the Semantic Test Suite.

If we add x to the “take” EP, we obtain a new dominance edge which makes the MRS into a net. Similar arguments apply for the two other sentences; for the tag question (1), one could even argue that the pronoun fragment shouldn’t even be in the semantic representation.

Although the STS is a very small corpus, it is designed to cover a range of semantic phenomena, and it is more reliably annotated with semantics than a random corpus of text. Hence we take these results as encouraging evidence that the Net Hypothesis is true.

4.3 A legitimate non-net

Nevertheless, we must mention that there is one type of MRS structures that seems to be linguistically plausible and still is a non-net.¹ One sentence of this type is

- (4) A woman the manager of whom fell ran.

A slightly simplified MRS for this sentence is shown in Fig. 4 on the left-hand side. This MRS is characterised by the fact that the two quantifiers bind variables in each other’s restrictions. It is not a net because there are two dominance edges from the root of the quantifier fragment a_y to the fragments $manager_{x,y}$ and run_y , but the only hypernormal path that connects these two fragments goes through the root of a_y , which is not allowed in a net. On the other hand, the MRS constitutes a

¹Thanks to Alex Lascarides and Ann Copestake for pointing out this example.

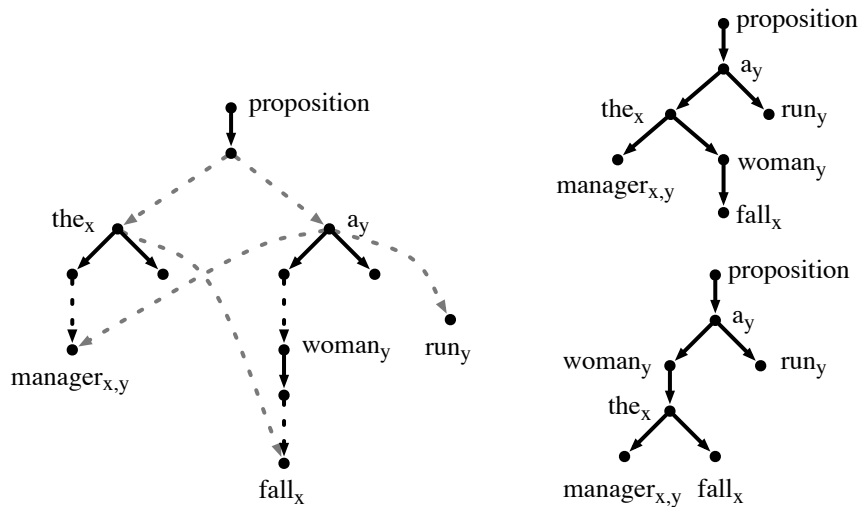


Figure 4: A linguistically legitimate non-net for the sentence “a woman the manager of whom fell ran” together with the two configurations permitted by the MRS.

plausible analysis for the sentence, as the two configurations (see Fig. 4) both are reasonable semantic representations.

However, examples of this type are extremely rare; we have not found a single MRS of this kind in the STS or the Redwoods or Rondane treebanks. In addition, the MRS in Fig. 4 can still be translated in a principled way into an equivalent normal dominance constraint which *is* a net and has the correct solved forms – although this translation is more complicated than the one used in the proof of Theorem 1. This means that there is probably a slight generalisation of nets for which the Net Hypothesis and the most important theorems about nets still hold.

5 Nets in Semantics Debugging

But now let’s assume that the Net Hypothesis is true – at least in a weaker form that says that *almost* all correct MRSs that are used in scope underspecification are nets. If this is the case, then the 17% non-nets that Fuchss et al. found must be due to errors in the syntax-semantics interface of the grammar. We can thus turn their finding around and use it to hunt for those rules in the grammar whose semantic components have bugs and which are responsible for generating the non-nets. In other words, we can use nets for semantics debugging.

5.1 Data Acquisition

The first step in this debugging process is to obtain a large collection of MRSs that are generated by the ERG. To this end, we repeated Fuchss et al.’s procedure of collecting MRSs for all parses of all sentences in the Verbmobil sections

| Treebank | Sentences | Parses | Ill-formed | Non-Nets | Nets |
|------------------|-----------|--------|------------|----------|--------|
| Verbmobil (VM6) | 2502 | 163814 | 33926 | 17921 | 111967 |
| Verbmobil (VM13) | 2093 | 159958 | 35634 | 20344 | 103980 |
| Verbmobil (VM31) | 1814 | 78332 | 11704 | 14504 | 52124 |
| Verbmobil (VM32) | 640 | 27017 | 3386 | 5280 | 18351 |
| Rondane | 805 | 38634 | 4381 | 5255 | 28998 |

Figure 5: Distribution of MRS structures for all parses of all sentences in the treebanks.

of the Redwoods 5 Treebank (Jan. 2005; 10503 sentences), and the Rondane Treebank (1034 sentences) distributed with the ERG, by parsing the sentences extracted from the treebank and extracting the MRS structures from the parses. We used the October 2004 version of the ERG. The numbers of sentences that could be parsed and the total numbers of parses (and therefore, MRSs) are shown in Fig. 5. This gave us a base number of almost half a million MRSs to work with.

We classified each sentence in the treebanks into one of three categories:

1. sentences whose MRS structure are not well-formed according to the shallow tests in the LKB system, such as structures containing free variables that aren't bound by any quantifier, or structures with cycles;
2. sentences whose MRS structures are well-formed according to the LKB checks, but are not nets, and
3. sentences whose MRS structures were nets.

In this way we collected about 63.000 non-nets.

The ratio of nets to non-nets among the well-formed MRS structures obtained from the Verbmobil corpora is 83 % to 17 %, so our results match those of Fuchss et al. (2004), which were based on a much smaller data set.

5.2 Semantic Debugging

In a second step, we then checked which rules are responsible for the introduction of the non-nets. We found that there are eleven rules which systematically derive only non-nets for all syntactic analyses of all sentences in the treebanks. These rules account for about 55% of the non-nets, and can be classified into four groups:

Measure noun phrases: MEASURE_NP, BARE_MEAS_NP

Coordinations of more than two conjuncts: P_COORD_MID, N_COORD_MID

Sentence fragments: FRAG_PP_S, FRAG_R_MOD_PP, FRAG_ADJ, FRAG_R_MOD_AP

Other rules: VPELLIPSIS_EXPL_LR, NUM_SEQ, TAGLR

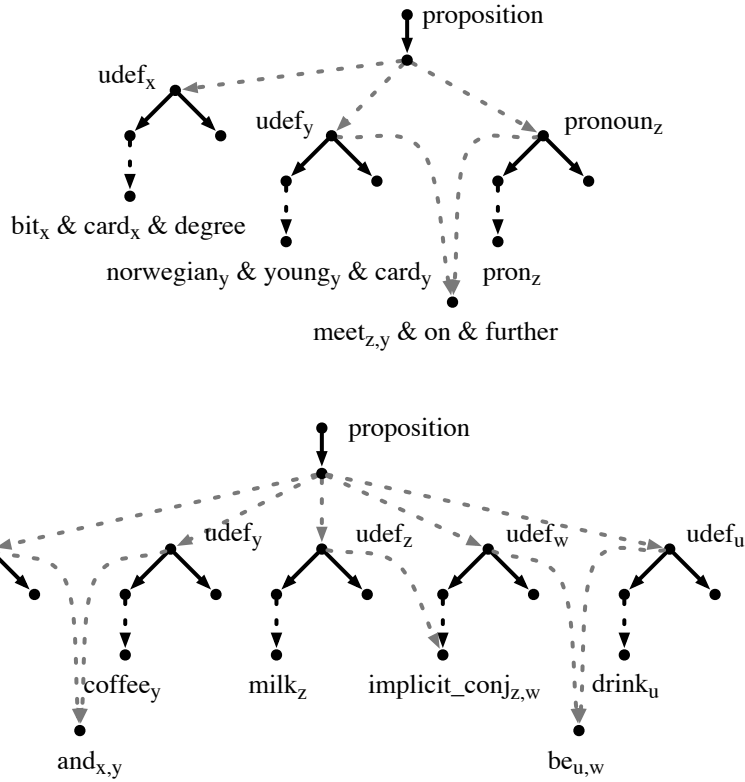


Figure 6: MRS structures for the annotated derivation for “a bit further on we meet two young Norwegians” (top) and “Drink is tea, milk and coffee” (bottom) in the Rondane treebank.

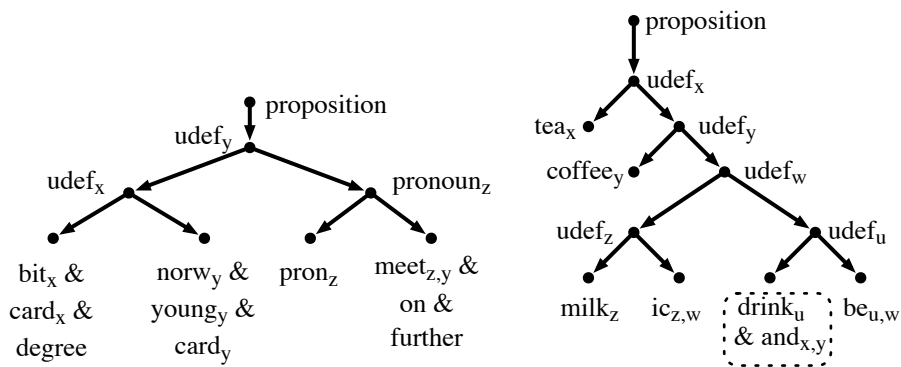


Figure 7: Configurations of the MRS structures in Fig. 6 that are meaningless as semantic representations.

We inspected these eleven rules by hand, and it turned out that indeed each of them had bugs in its semantics component. Typical bugs are that the MRSs they generate either have too few occurrences of a bound variable (which leads to missing dominance edges) or that EPs that should form a single fragment (e.g. by EP conjunction) are split into separate fragments.

Consider, by way of illustration, the two MRS structures shown in Fig. 6. The first MRS is derived by the ERG for the sentence “A bit further on we meet two young Norwegians” (Rondane 996). In this MRS, the quantifier “a bit,” whose analysis uses the MEASURE_NP rule, introduces a bound variable x that is used only in its restriction, but in none of the predicates in its scope (“meet further on”). This is obviously not intended. Because the missing variable binding also relaxes the constraints on how fragments can be plugged together, the underspecified description admits structurally wrong readings, e.g. by plugging “young Norwegian” into the scope of “a bit” (see Fig. 7). If we fix the structure by using x in the EPs for “further on”, this introduces an additional dominance edge in the graph which makes the structure a net.

A similar bug occurs in the second MRS structure, which has been derived from the sentence “Drink is tea, milk, and coffee” (Rondane 1412) by using the N_COORD_MID rule. The EPs “and” and “implicit_conj” are two different components of the same collective “tea, milk, and coffee”, and should therefore be connected. Because they aren’t, the structure has meaningless configurations such as the one shown in Fig. 7, in which “and” and “drink” have been merged into the same argument handle (and almost 1000 further configurations). If we connect “and” and “drink” either by combining them into a single EP-conjunction or by introducing additional material (e.g., a quantifier fragment) that connects the two nodes, the MRS structure again becomes a net.

A further example is the graph at the top left in Fig. 3, whose derivation uses the TAGLR rule.

5.3 Discussion and Analysis

To summarise, our analysis of the rules that generate non-nets pointed us towards a list of eleven rules, each of which contained bugs. What’s more, each of these rules generates *well-formed* MRS structures. This means that they could never have been found using the shallower checks that the LKB system already offers. Hence the Net Hypothesis is true enough to be useful for finding grammar rules with erroneous semantic components.

We corrected some of the rules by hand; these corrections are already part of the ERG version of February 05. The correction of the other rules is ongoing work. In order to measure the progress that this makes, we compared the number of ill-formed MRSs and non-nets before and after the correction – this time only on the parses that were actually annotated in the Rondane treebank. The result of this evaluation is shown in Fig. 8. Because the treebank contains only annotations for sentences that can be analysed by the underlying grammar, the two versions

| Treebank | Sentences | Ill-formed | Non-Nets | Nets |
|-------------------------|-----------|------------|----------|--------|
| Rondane (October 2004) | 1034 | 7.5 % | 11.1 % | 81.4 % |
| — | 942 | 6.8 % | 10.5 % | 82.7 % |
| Rondane (February 2005) | 961 | 2.5 % | 7.9 % | 89.6 % |
| — | 942 | 2.4 % | 8 % | 89.6 % |

Figure 8: Classification of the sentences in the Rondane treebank for the original and the partially corrected version of the ERG.

of the treebank contain slightly different sets of derivation trees. To allow for a proper comparison, we report the results both for each complete treebank and for the subset of sentences that is present in both treebanks.

It turns out that the percentage of ill-formed MRSs in Rondane has dropped considerably, which is a clear indicator that the overall correctness of the MRSs has improved. In addition, the percentage of non-nets has also gone down significantly. If we only count well-formed MRSs, 92% of the MRSs in the corrected treebank are nets, which we take as further support for the Net Hypothesis.

6 Conclusion

We have shown that nets can be a useful tool for debugging the semantics component of a large-scale grammar. All eleven rules in the ERG that computed only non-nets turned out to be semantically problematic; a typical error was that a bound variable was not used where it should be. None of these rules could have been found easily by the existing well-formedness tests in the LKB system.

In addition, we have presented further support for the Net Hypothesis. Only three of the 107 MRS structures in the Semantics Test Suite are non-nets, and we have argued that these three MRSs are indeed missing dominance edges. Also, the partially corrected ERG derives about 90% nets on the Rondane treebank. Nevertheless, there are also (rare) MRS structures that seem to be legitimate non-nets. Generalising the definition of a net to encompass these MRSs is an important issue for future research.

The concept of a net seems to be rather complicated at first glance. However, there are portable and efficient tools for checking whether an MRS structure is a net. Utool, the Swiss Army Knife of Underspecification (Koller and Thater, 2005), can be used to solve underspecified descriptions and also implements a linear-time net test, and supports MRS as an input formalism. This tool takes less than half an hour to check which MRSs for all parses of the sentences in the Rondane treebank are nets, i.e. each MRS takes about fifty milliseconds on average. Utool is available from <http://utool.sourceforge.net>.

There are various further ways in which the work we report here could be extended. On the one hand, it would be interesting to see whether a similar debugging methodology would yield problem rules based on the LKB’s well-formedness tests,

and it would be natural to look not just for problematic *rules*, but also for problematic *lexicon entries* this way. On the other hand, we have only used a very coarse heuristic in finding the rules that are responsible for the generation of the non-nets. We suspect that some semantically problematic MRS structures are derived not by a single rule, but by a combination of rules. One way of finding such rule combinations would be to analyse the MRSs for a corpus with a decision tree learner, which would try to derive rules that capture such combinations.

References

- Bender, Emily M., Flickinger, Dan and Oepen, Stephan. 2002. The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. In J. Carroll, N. Oostdijk and R. Sutcliffe (eds.), *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th COLING*, Taipei, Taiwan.
- Bodirsky, M., Duchier, D., Niehren, J. and Miele, S. 2004. A New Algorithm for Normal Dominance Constraints. In *Proc. SODA*.
- Bos, J. 1996. Predicate Logic Unplugged. In *Proc. 10th Amsterdam Colloquium*, pages 133–143.
- Copestake, A. and Flickinger, D. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proc. LREC*.
- Copestake, A., Flickinger, D., Pollard, C. and Sag, I. 2004. Minimal Recursion Semantics: An Introduction. *Journal of Language and Computation* To appear.
- Copestake, A., Lascarides, A. and Flickinger, D. 2001. An Algebra for Semantic Construction in Constraint-based Grammars. In *Proc. 39th ACL*, Toulouse.
- Egg, M., Koller, A. and Niehren, J. 2001. The Constraint Language for Lambda Structures. *Journal of Logic, Language, and Information* 10, 457–485.
- Fuchss, R., Koller, A., Niehren, J. and Thater, S. 2004. Minimal Recursion Semantics as Dominance Constraints: Translation, Evaluation, and Analysis. In *Proc. 42nd ACL*, Barcelona.
- Koller, Alexander and Thater, Stefan. 2005. Efficient solving and exploration of scope ambiguities. In *Proceedings of the ACL-05 Demo Session*.
- Niehren, J. and Thater, S. 2003. Bridging the Gap between Underspecification Formalisms: Minimal Recursion Semantics as Dominance Constraints. In *Proc. 41st ACL*, Sapporo.
- Oepen, S., Toutanova, K., Shieber, S., Manning, C., Flickinger, D. and Brants, T. 2002. The LinGO Redwoods Treebank: Motivation and Preliminary Applications. In *Proc. 19th COLING*.