Mapping between Dependency Structures and Compositional Semantic Representations

Max Jakob[1,3], **Markéta Lopatková**[1], **Valia Kordoni**[2,3]

[1]Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic
[2]German Research Centre for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany
[3]Department of Computational Linguistics, Saarland University, Saarbrücken, Germany
maxj@coli.uni-sb.de, lopatkova@ufal.mff.cuni.cz, kordoni@dfki.de

Abstract

This paper investigates the mapping between two semantic formalisms, namely the tectogrammatical layer of the Prague Dependency Treebank 2.0 (PDT) and (Robust) Minimal Recursion Semantics ((R)MRS). It is a first attempt to relate the dependency-based annotation scheme of PDT to a compositional semantics approach like (R)MRS. A mapping algorithm that converts PDT trees to (R)MRS structures is developed, associating (R)MRSs to each node on the dependency tree. Furthermore, composition rules are formulated and the relation between dependency in PDT and semantic heads in (R)MRS is analyzed. It turns out that structure and dependencies, morphological categories and some coreferences can be preserved in the target structures. Moreover, valency and free modifications are distinguished using the valency dictionary of PDT as an additional resource. The validation results show that systematically correct underspecified target representations can be obtained by a rule-based mapping approach, which is an indicator that (R)MRS is indeed robust in relation to the formal representation of Czech data. This finding is novel, as Czech, with its free word order and rich morphology, is typologically different than languages analyzed with (R)MRS to date.

1. Introduction

This paper introduces a method for mapping between dependency structures and compositional semantic representations. Recent studies of natural language processing have shown a clear and steady shift of focus from pure syntactic analyses to more semantically informed structures. As a result, we have seen an emerging interest in parser evaluation, for instance, based on more theory-neutral and semantically informed representations, such as dependency structures, like the ones of the Prague Dependency Treebank (PDT; Hajič (2006)). Moreover, various existing parsing systems are nowadays also adapted to provide semantic information in their outputs. The obvious advantage in such an approach is that one can derive more fine-grained representations which are not typically available from shallow semantic parsers (e.g., modality and negation, quantifiers and scopes, etc.). To this effect, various semantic representations have been proposed and used in different parsing systems. One of them is Robust Minimal Recursion Semantics (RMRS; Copestake (2007a)), which we focus on in the work reported in this paper. Generally speaking, semantic representations like the ones of RMRS should be capable of embedding dependency structure information, which is so important nowadays for parser evaluation. However, it is proven to be non-trivial to map even basic dependency information among different compositional semantic representations. This becomes a barrier to both sides, i.e., to compositional semantic based approaches and to dependency structure approaches, making the cross-fertilization of systems and resources using different semantic representations very difficult.

In this paper, we present an approach towards mapping between dependency and compositional semantic representations. More specifically, we map Prague Dependency Treebank dependency annotations to Robust Minimal Recursion Semantics structures. Evaluation results show that the mapping from PDT to RMRS structures is reliable and beneficial to deep parsing in general. Finally, the shown mapping of a dependency based annotation onto RMRS makes the PDT and its backbone theoretical linguistic formalism available to a bigger community of NLP researchers. This endeavor is also novel in that it explores the capability of RMRS to represent typologically different languages, in this case, a free word order language with a rich morphology, like Czech.

Other works address similar problems of mapping between different formalisms, involving either RMRS or PDT. These works include the projection of German dependencies to RMRS (Spreyer and Frank, 2005) and the mapping of generic logical forms in frame-like notation onto Minimal Recursion Semantics structures (Allen et al., 2007). The reverse approach is taken by Žabokrtský et al. (2008) in that they try to represent language data other than Czech in the PDT annotation scheme.

2. Source and target representations

2.1. Dependency Representation

The Prague Dependency Treebank 2.0 (Hajič et al., 2006) is an annotated corpus of Czech data. It is based on the long-standing tradition of the Prague Linguistic Circle and was adapted for the current computational linguistics research needs. PDT adopts a stratificational, dependency-based approach to language representation. Three layers are distinguished, ranging from morphology through syntax to semantics and beyond. The layer of maximal abstraction is represented by the tectogrammatical layer, annotating semantic structure and dependencies (i.e., semantic head-dependent relation), topic-focus articulation, coreferences and morphological categories. The information of this layer, captured in a form of dependency tree struc-

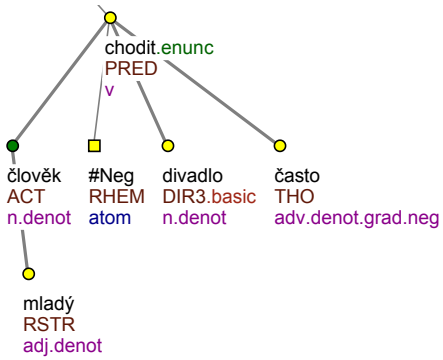

Figure 1: Tectogrammatical tree for the sentence string "*Mladí lidé nechodí do divadel často.*" (English "*Young people don't go to the theater frequently.*")

ture – its nodes represent individual lexical nodes and their morphological categories and edges represent syntactic relations – serves as an input for the developed mapping. An example tree is shown in figure 1. Furthermore, knowledge about valency for Czech words contained in the valency dictionary of the PDT (Hajič et al., 2003) is incorporated in the target representation.

2.2. Compositional Semantic Representation

Because of its underspecifiability, Minimal Recursion Semantics (MRS; Copestake et al. (2005)) has been widely used in many deep and shallow processing systems. The main assumption behind MRS is that the interesting linguistic units for computational semantics are the *elementary predications* (EPs), which are single relations with associated arguments. Robust Minimal Recursion Semantics (RMRS; (Copestake, 2007a)) is a variant of the MRS formalism that attempts to formalize a semantic description that can be used by both deep and shallow processing techniques. Therefore, it is possible in RMRS to underspecify and to dynamically add predicate arguments.

RMRS is the target representation of the mapping presented here. An RMRS structure consists of a quadruple ⟨ hook, EP bag, arguments set, handle constraints ⟩[1] (Copestake, 2007b). The first element is the *hook* of the structure. It is important during the semantic composition of complex RMRSs. The second element is the *EP bag*. It is a set of predicates that describes the lexical and some relational semantic information contained in a sentence. The third element is a set of arguments, that belong to the individual elements of the EP bag. The last element is a set of *handle constraints* that specify certain scopal relations of the elements in the EP bag.

(R)MRS itself is a formalism, not a semantic theory (Copestake et al., 2005). PDT adopts the Functional Generative Description (FGD; Sgall et al. (1986)) as its theoretical background. The task is to adapt the abstraction level and annotation features of PDT and to attempt to represent them in the RMRS formalism.

[1]Variable equalities are directly resolved in this project.

3. Mapping

We introduce a practical task to map the dependency-based annotation scheme of the tectogrammatical layer of PDT onto the RMRS formalism. The target representation incorporates knowledge about tree structure and dependency relations, morphological categories, as well as certain grammatical coreferences (control structures, complex predicates and some types of reciprocity). Other tectogrammatical information, such as textual coreference and topic-focus articulation, has no correspondence in the target representation yet. Furthermore, certain constructions are skipped due to their linguistic complexity that exceeded the range of this basic research.

The developed approach constructs RMRS structures for each node of a tectogrammatical tree. These so called *node-RMRS*s represent the semantics of the subtrees rooted at the respective nodes. The node-RMRS of the root node captures the semantics of the complete sentence. A node-RMRS is created in two steps. The first step initializes the lexical and morphological information, using the attributes of the current node. The second step assembles more complex structures by combining multiple node-RMRSs. It makes use of the explicit dependency relation in the tectogrammatical tree, as well as to a special relation in the case of valency modifications in coordinations (see section 3.3.).

3.1. node-RMRS Initialization

The core element of each node-RMRS is the *lexical* EP. It is created and added to the EP bag during the initialization step. Lexical EPs represent the lexical units of the sentence, carrying the tectogrammatical lemma of nodes, its semantic part-of-speech and potentially the index of a valency frame (as additional sense distinction) in the relation name. Morphological categories are mapped to features of variables that are introduced for ARG0 (the first argument of an EP). In English (R)MRS structures, this is typical done for person, number and gender, as well as tense, aspect and mood. For Czech, we need to extend the number of these features due to its rich morphology.

The elements of the hook feature of a node-RMRS, namely top label, top anchor and index variable (Copestake, 2007b), are initialized to the label, anchor and ARG0 of the lexical EP. Quantifiers are introduced for nominal objects, including the typical constraint for the restriction hole. The body remains unconstrained. Examples (1) and (2) show the initialized node-RMRSs for the nodes for *chodit* ('*to go*') and *člověk* ('*man*') in figure 2. There is a conflict in the typical naming of quantifier restrictions and the RSTR functor in PDT. To guarantee that arguments and EPs named RSTR always carry the meaning of the functor, we were forced to rename the quantifier restriction in this project.

(1) $< [l1, a1, e1],$
$\{l1{:}a1{:}\text{_chodit_v_1}(e1_{[\ldots,tense:sim,verbmod:ind]})\},$
$\{\ \},$
$\{\ \} >$

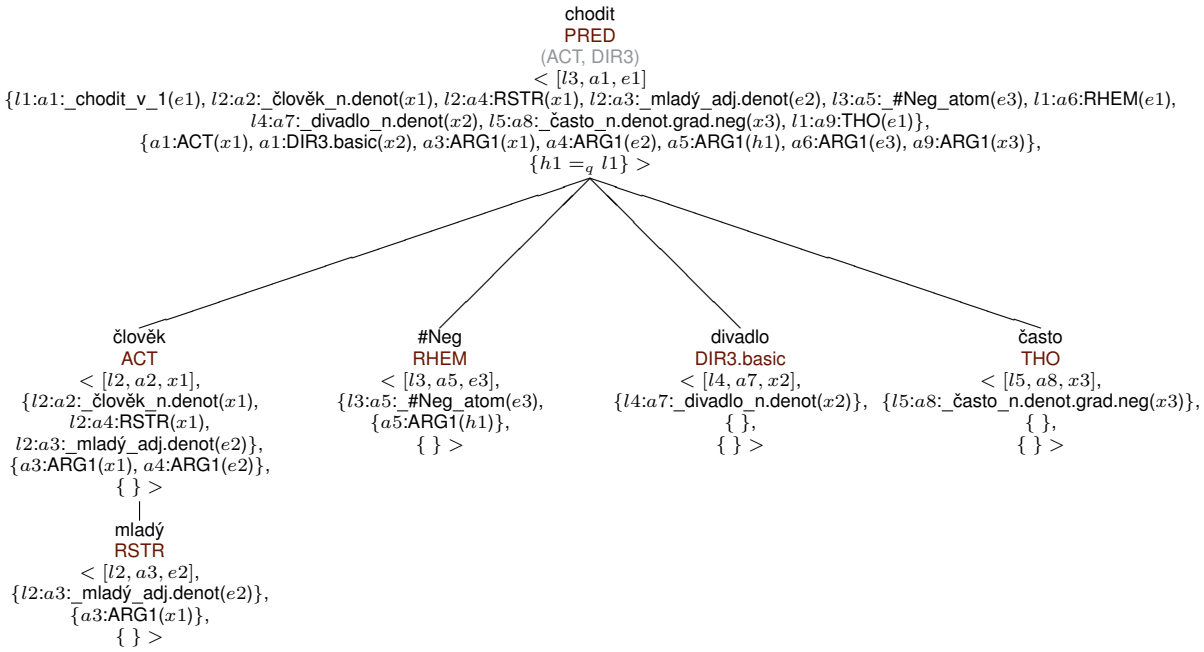

Figure 2: Tree of figure 1 including node-RMRSs. Valency positions of the frame-bearing words are shown as a set of functors in round brackets. Quantifiers and variable features are omitted for clarity.

(2) $< [l2, a2, x1],$
$\{l2{:}a2{:}\text{_člověk_n.denot}(x1_{[gender:anim,number:pl]}),$
$l6{:}a10{:}\text{udef_q}(x1)\},$
$\{a10{:}\text{RESTRICTION}(h2), a10{:}\text{BODY}(h3)\},$
$\{h2 =_q l2\} >$

For nouns governing adjectives, the same nominal variable must be used (as in *člověk*, '*man*' and *mladý*, '*young*'). Therefore, we re-use parts of the governing hook information in the dependent node-RMRS. The composition of more complex node-RMRSs is outlined next.

3.2. node-RMRS Composition

The initialized node-RMRSs, containing the lexical EPs, are combined to build more complex structures. This procedure includes the adding of additional arguments, EPs and constraints. The concrete details depend on the generalized semantic part-of-speech of the involved nodes. Furthermore, we distinguish valency modification, optional free modification and coordination. Most importantly, in this composition step, the EP bag, argument set and handle constraints of the dependent node-RMRS are joined together with those of the governing node-RMRS.

3.2.1. Valency Modification

Certain tectogrammatical nodes, namely verbs and some types of nouns and adjectives, have annotated valency frames. These frames are registered in the PDT valency lexicon (Hajič et al., 2003) that stores all obligatory and optional valency modifications for the above mentioned Czech words. Each valency frame consists of several valency slots, characterized by a functor, i.e., label for a type of valency modification. Using this information, it is possible to determine valency modifications in tectogrammatical trees.

Valency modifications are labeled with six functors: ACT, PAT, ADDR, EFF, ORIG and MAT. Theoretically, the whole set of valency modifications must be registered in the PDT valency lexicon. It may, however, be the case that one of these functors occurs in the data without being present in the respective governing valency frame. This incomplete annotation can, in this case, be overcome by the capability of RMRS to dynamically add arguments to EPs. The named functors are always correctly treated as valency modifications. In addition to valency modifications, valency frames also store information on obligatory free modifications. They are treated exactly in the same way as valency modifications and together make up the arguments of EPs.

Depending on the semantic part-of-speech of the dependent node, valency modification works in different ways. In fact, four types ought to be distinguished based on the generalized semantic part-of-speech of either *n*, *adj*, *v*, or *adv*. In general, an argument relation is created that has the dependent functor as its name. Moreover, it is attached to the lexical EP of the governing node-RMRS by anchor equality and it contains the index variable of the dependent node-RMRS (verbs being an exception). This argument relation is added to the arguments set of the governing node-RMRS. Figures 3, 4 and 5 illustrate examples of the behavior of valency modifications with different parts-of-speech at the dependent nodes. In figure 1, the modifications with the ACT and the DIR1.basic functor are both valency modifications.

3.2.2. Optional Free Modification

Non-valency modifications are considered to be free modifications. If they are obligatory and therefore listed in a nodes' valency frame (like DIR3 in figure 2), they are treated in the same way as valency modifications. Otherwise, a new EP is introduced that has the functor of the dependent node as relation name. This EP is in an implicit conjunction with the governing node's lexical EP, from which it also copies the argument in ARG0. Variables

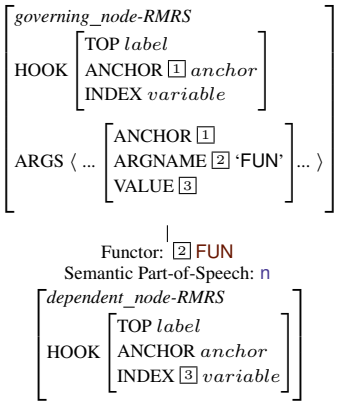

Figure 3: Valency composition for nouns as dependent

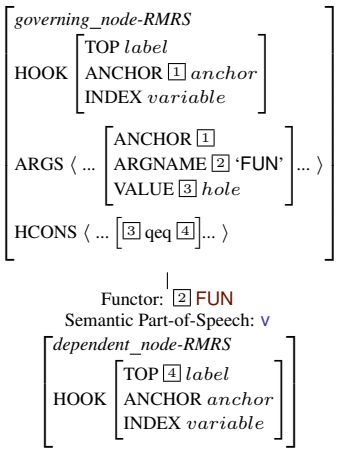

Figure 4: Valency composition for verbs as dependent

and handle constraints are constructed in a similar way to valency modifications, except that the variables are added to ARG1 of this newly introduced EP instead of the argument relation for the lexical EP (figure 6). In figure 2, the modification with the THO functor is an optional free modification. The functor information is preserved through the newly established EP.

3.2.3. Coordination

Coordination and apposition structures are represented by a special type of nodes in the dependency trees. They introduce tree edges capturing non-dependency relations that merely group nodes together. Dependent nodes of coordinations that are annotated as being members of the coordination are linked together in the target formalism using a chain of binary relations following the original approach of Copestake et al. (2005), although an n-ary EP would also be possible. Figures 7 and 8 present an example. If the coordination node is processed as a dependent in the composition of another node-RMRS, it inherits the functor and the generalized semantic part-of-speech from its member nodes.

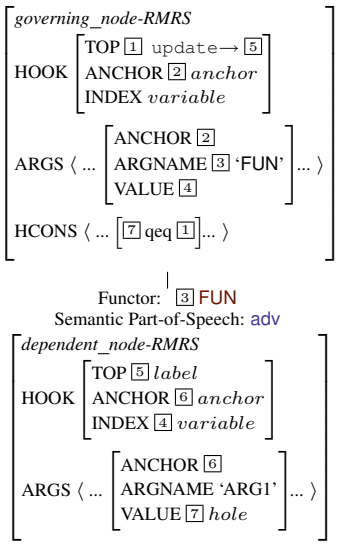

Figure 5: Valency composition for adverbs and other scopal elements as dependent

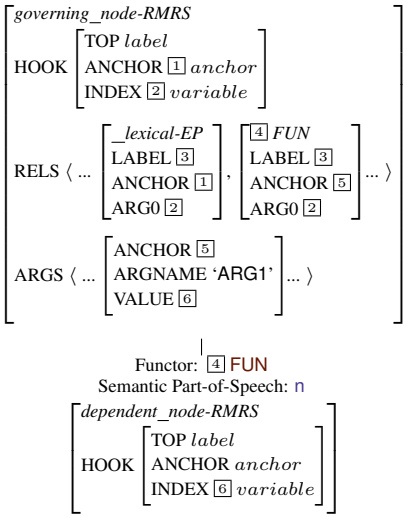

Figure 6: Composition for optional free modification with a noun

3.3. Dependent Nodes for Composition

For input trees containing no coordination or apposition nodes, the two steps just presented can be applied recursively to each node (initialization) and to all its direct dependent nodes (composition). On the other hand, coordination and apposition nodes complicate the issue of determining the appropriate set of dependents for the composition step. Figure 8 shows a tectogrammatical tree containing a coordination node (*a*, '*and*'). This coordination has two member nodes (*řídit*, '*to command*' and *kontrolovat*, '*to control*'), marked by _M. The non-member nodes modify all members of the coordination. Specifically, the node for *StB* ('*State Security*') fills the actor valency slot of both verbs being member nodes and the node for *skupina* ('*group*') fills both patient valency slots.

As described above, valency modifications fill their slot in

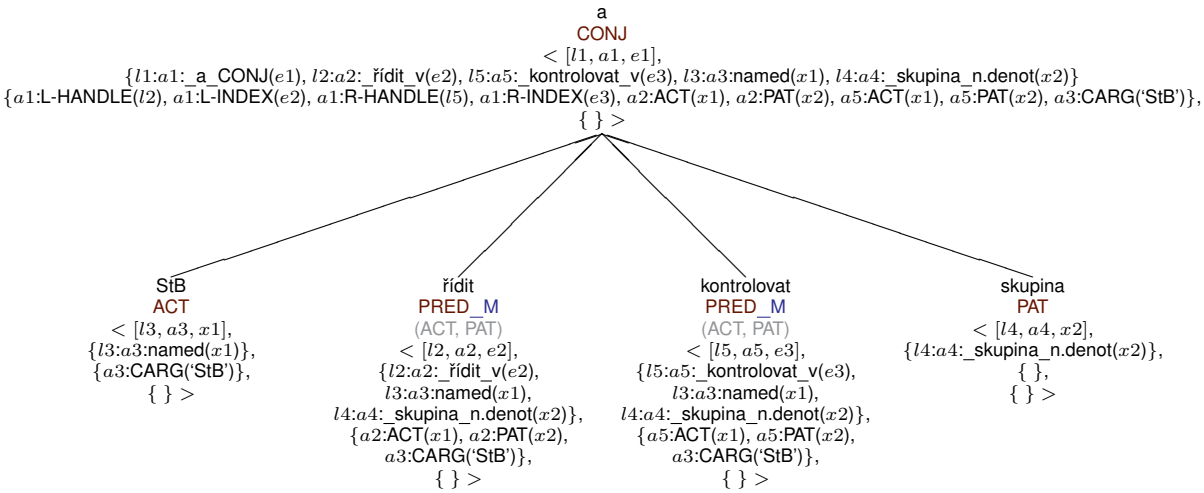

Figure 7: Tree of figure 8 including node-RMRSs. Valency positions of the frame-bearing words are shown as a set of functors in round brackets. Quantifiers and variable features are omitted for clarity.

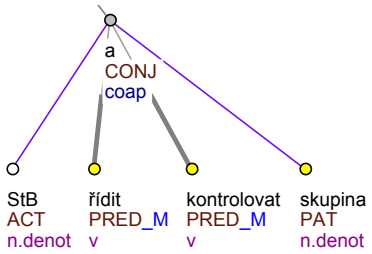

Figure 8: Tectogrammatical tree for the sentence string "*StB řídila a kontrolovala skupinu.*" (English "*The StB* (State Security) *commanded and controlled the group.*")

the governing lexical EP by creation of a new argument relation. In figure 8, the node for *StB* must therefore be processed in the composition step of both node-RMRSs for the *řídit* node and the *kontrolovat* node. That means we need to refine the method of determining the dependent nodes that are relevant for the composition step. Additionally to the direct dependents, valency modifications that are not in the subtree under the current node have to be included as well (but excluded when processing the coordination node). Furthermore, if these kind of valency modifications are themselves joined under a coordination (cf., the sentence string of figure 8 starting with "*The StB and others commanded* ..."), the coordination node-RMRS has to be processed for composition for the simple fact that there is only a single slot to be filled. Lastly, if the valency modification in the described position is an adverb (a scopal EP), the handle constraint must be introduced to outscope the coordination node. Otherwise, both member nodes' top would be updated to the same label, which would result in the coordination node not linking distinct structures.

3.4. Control Structures

In control structures (and similarly in complex predicates and some types of reciprocity) in PDT, one node points to another node in the same sentence and thus inherit some of its features. This is annotated using coreference links. For RMRS structures, shared nominal variables in two different verbal EPs is the correct treatment of this phenomenon. In order to provide this variable identity, the hook of the node with the coreference link is inherited from the antecedent node that the grammatical coreference points to. Figures 9 and 10 present an example. The *#Cor* node inherits the hook of the node-RMRS for *pozornost* ('*attention*'), resulting in variable $x1$ being shared by the actor slots of both verbs *dokázat* ('*can*') and *vytvořit* ('*to create*').

3.5. Algorithm

The general procedure of the mapping algorithm is shown as algorithm 1. The node-RMRSs are constructed in a top-down recursive approach, storing previously computed structures in a table to avoid re-computations. First, coreference links are resolved (steps 1-3) to deal with structures described in the last section. The lexical EPs, and generalized quantifiers if appropriate, are introduced in the initialization (step 4). The relevant nodes for composing the RMRS for the subtree rooted at the current node are calculated with respect to the complex interplay between dependency, coordination edges and valency (step 5). The cycle (steps 6-10) composes more complex node-RMRSs by means of valency modification, optional free modification or coordination and propagates all EPs, arguments and constraints up to the higher tree level. The last step (11) adds elided optional valency arguments to the current lexical EP, which is important for the interconversion between MRS and RMRS. To obtain the RMRS structure for the complete tectogrammatical tree, the *get_node-RMRS* function is called with the root node. Several details are omitted for simplicity.

4. Evaluation

The source dependency trees are represented accurately in RMRS regarding their structure and dependencies, morphological categories and some types of coreferences. But the built RMRS structures differ from the outputs that are typically produced from HPSG parsers. Which input tokens introduce EPs, for instance, depends on whether the

Algorithm 1 *get_node-RMRS* (general procedure)

Input: tectogrammatical node
Output: RMRS structure

```
1  if coreference annotated control structure or the like then
2      current node ← grammatical antecedent node
3  end if
4  initialize the node-RMRS for the current node
5  obtain dependent nodes relevant for RMRS composition
6  for all these relevant nodes in pre-defined order do
7      dependent_node-RMRS ← get_node-RMRS(node)
8      treat dependent_node-RMRS as
       a member of a coordination or
       a valency modification or
       a free modification
9      merge all information with the governing node-RMRS
10 end for
11 add empty argument positions for unfilled valency slots
12 return node-RMRS
```

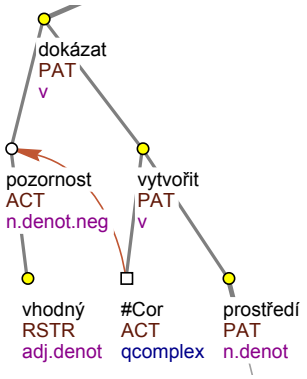

Figure 9: Tectogrammatical subtree for the substring "[...] *vhodná pozornost dokáže vytvořit prostředí* [...]" (English "[...] *appropriate attention can create an environment of* [...]")

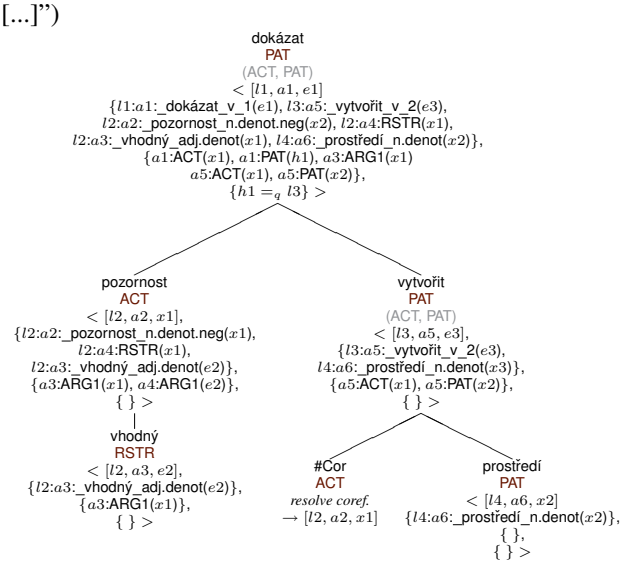

Figure 10: Subtree of figure 9 including node-RMRSs. Valency functors are shown in round brackets. Quantifiers and variable features are omitted for clarity.

token has a correspondence on the tectogrammatical layer. Following FGD, function words (including prepositions), for example, are not represented by nodes but by functors and other node attributes. Functor information is preserved in argument relations (valency modification) and in additional EPs (optional free modification). Further, the number of features on variables is higher than for English RMRSs, incorporating all morphological categories annotated at the tree nodes. However, where possible, we tried to be close to the behavior of the English Resource Grammar (Copestake and Flickinger, 2000). Note that dominance of adverbs, negation, etc., is represented in reverse in the target structures, but the scope remains underspecified in both representations.

After this basic research, there are still some open issues. Specific generalized quantifiers were not implemented yet (Czech correspondences of *all*, *some*, etc.). It would require information from the lower annotation layers or a generation step, since quantifiers and negative pronouns (*all* vs. *any*) are distinguished by morphological categories only. For now, all nominal variables are bound by an underspecified generalized quantifier EP (as shown in example (2)). The mapping of character positions of the word tokens was omitted here but will be incorporated in order to capture word order. The attempt to represent coordinations whose member nodes were inconsistent regarding functor or generalized part-of-speech information failed. Trees containing this kind of structures were skipped, as were coordinations of adverbs, idioms and some types of comparative constructions. Textual coreferences and topic-focus articulation were not mapped.

4.1. Validation

The automatic validation involved two structural criteria. As a first step, the RMRS structures were converted to MRS representations. This is feasible since unspecified valency modifications were made explicit. Then, the *net criterion* (Flickinger et al., 2005) was tested. It roughly says that the only relevant structures in practice consist of fragments that obey two structural properties. They must not contain open hole and/or ill-formed island fragments (see below). On a positive outcome, additionally, it was tested if the MRS had any configurations, i.e., linguistic readings. Only if both these criteria were met, the MRS was regarded to be structurally valid. Given this characterization of valid structures, the recall for the mapping of the complete PDT corpus is 80.93 % (see table 1).

Structures that failed the test can be put in three classes, as shown in table 2. These error classes are named *no configuration*, *open-holes* and *ill-formed islands* . Source trees in all of these classes showed common characteristics. The majority of structures for which no configuration exists are produced from source trees that violate the assumption that coreference links for control structures point to nouns. This results in unresolvable handle constraints. Open-hole fragments lack EPs to be embedded by a quantifier's body. In our case, this was caused exclusively by source trees with independent nominative clauses (DENOM functor) or par-

mapped	44580	90.19 %
skipped	4851	9.81 %
valid	40004	80.93 %
Precision	40004/44580	89.74 %
Recall	40004/49431	80.93 %

Table 1: Precision and Recall for structurally valid MRSs

Nets	have configuration	**40004**
	no configuration	24
Non-Nets	open hole	4539
	ill-formed island	13
	mapped	44580

Table 2: Result totals of produced MRSs

enthetic clauses (PAR functor) at the root. The lack of a main verb explains their structural incompleteness. Ill-formed island fragments lack a certain constraint structure. However, Flickinger et al. (2005) admit that – among ill-formed island structures – there are special cases that make them "legitemate non-nets". The ill-formed islands produced in this mapping all fall into this case. The source trees all have coreferences links from a node to one of its own ancestor nodes. The last two observations put the invalidity of the non-net structures in question.

The overall performance is regarded to be very satisfactory. It gives rise to the notion that the two involved formalisms are basically compatible and the mapping is feasible. Current efforts are being made to investigate the re-mapping of the RMRS structures to PDT trees, which would enable the creation of PDT-style trees from the big amount of available RMRS data. Furthermore, there is still a number of attributes and phenomena that is lost or skipped during the mapping. We plan to further extend the research.

5. Conclusion

In this paper, we have presented a study of mapping between PDT-like dependency structures and semantic structures in the form of (R)MRS. The experiment results show that there can indeed be a reliable mapping from dependency structures to compositional semantic representations. It is also important to realize that such kind of mapping will also enable us to enrich deep (HPSG) syntactic and semantic annotations with fine-grained dependencies from valuable manually created resources, like PDT. The outcome enriched semantic structures can eventually be helpful for applications like question answering and information extraction, for instance, in the future.

6. Acknowledgements

The authors acknowledge support from the Erasmus Mundus Program through the European Masters Program in Language and Communication Technologies (LCT). The work on this project was also partially supported by the grant of MŠMT ČR No. MSM0021620838.

7. References

James Allen, Myroslava Dzikovsk, Mehdi Manshadi, and Mary Swift. 2007. Deep Linguistic Processing for Spoken Dialogue Systems. In *Proceedings of the ACL Workshop on Deep Linguistic Processing*, pages 49–56, Prague, Czech Republic, June.

Ann Copestake and Dan Flickinger. 2000. Open source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 591–598, Athens, Greece.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal Recursion Semantics: An Introduction. *Research on Language and Computation*, 3(4):281–332.

Ann Copestake. 2007a. Applying Robust Semantics. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 1–12, Melbourne, Australia.

Ann Copestake. 2007b. Semantic composition with (Robust) Minimal Recursion Semantics. In *Proceedings of the ACL Workshop on Deep Linguistic Processing*, pages 73–80, Prague, Czech Republic, June.

Dan Flickinger, Alexander Koller, and Stefan Thater. 2005. A new well-formedness criterion for semantics debugging. In Stefan Müller, editor, *Proceedings of the 12th International Conference on HPSG*.

Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. 2003. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68. Vaxjo University Press.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková-Razímová. 2006. Prague Dependency Treebank 2.0. http://ufal.mff.cuni.cz/pdt2.0/.

Jan Hajič. 2006. Complex Corpus Annotation: The Prague Dependency Treebank. In Mária Šimková, editor, *Insight into the Slovak and Czech Corpus Linguistics*. Veda, Bratislava, Slovakia.

Petr Sgall, Eva Hajičová, and Jarmila Panenová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing, Prague, Czech Republic/Dordrecht, Netherlands.

Kathrin Spreyer and Anette Frank. 2005. Projecting RMRS from TIGER Dependencies. In Stefan Müller, editor, *Proceedings of the 12th HPSG Conference*, pages 354–363, Lisbon, Portugal.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. Tecto MT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, USA. Association for Computational Linguistics.