Conversion of a Russian dependency treebank into HPSG derivations

Tania Avgustinova and Yi Zhang

Language Technology Lab
DFKI GmbH

{avgustinova; yzhang} @ dfki.de

Abstract

The Russian syntactic treebank SynTagRus is annotated with dependency structures in line with the Meaning-Text Theory (MTT). In order to benefit from the detailed syntactic annotation in SynTagRus and facilitate the development of a Russian Resource Grammar (RRG) in the framework of Head-driven Phrase Structure Grammar (HPSG), we need to convert the dependency structures into HPSG derivation trees. Our pilot study has shown that many of the constructions can be converted systematically with simple rules. In order to extend the depth and coverage of this conversion, we need to implement conversion heuristics that produce linguistically sound HPSG derivations. As a result we obtain a structured set of correspondences between MTT surface syntactic relations and HPSG phrasal types, which enable the cross-theoretical transfer of insightful syntactic analyses and formalized deep linguistic knowledge. The converted treebank SynTagRus++ is annotated with HPSG structures and of crucial importance to the RRG under development, as our goal is to ensure an optimal and efficient grammar engineering cycle through dynamic coupling of the treebank and the grammar.

1 Introduction

Key issues brought up recently in the research and development community concern the application of treebanks in acquiring linguistic knowledge for natural language processing, the role of linguistic theories in treebank development, and the suitability of treebanks as a basis for linguistic research.[*] In this context, we discuss the conversion of a Russian dependency treebank into Head-driven Phrase Structure Grammar (HPSG) derivations needed in the context of the Russian Resource Grammar (RRG) under development in our group ([3], [4]). We shall, therefore, focus on the problems of annotation transfer revealing possibilities for conceptual alignment of the underlying linguistic theories. Other aspects that will be

[*] We are grateful to Leonid L. Iomdin for providing us with access to the SynTagRus dependency treebank and for helpful answers to annotation-related questions.

touched upon are related to the use of a bootstrapping approach towards an incremental treebank conversion process.

The Russian treebank SynTagRus – cf. [6], [7], [2] – contains a genuine dependency annotation theoretically grounded in the long tradition of dependency grammar represented by the work of Tesnière [15] and Mel'čuk [10] among others. In particular, a complete dependency tree is provided for every sentence in the corpus. Supplied with comprehensive linguistic annotation, this treebank has already served as a basis for experimental investigations using data-driven methods [13]. By way of background, we start by introducing the Meaning-Text Theory (MTT) tradition of dependency grammar as reflected in the syntactic annotation of the Russian treebank SynTagRus and obtained with the ETAP-3 linguistic processor [1]. The main part of the paper is then devoted to the step-by-step "on-demand" conversion of the original dependency representation into an HPSG-conform phrase structure format. Finally, we discuss some non-trivial cross-theoretical issues and consider possibilities for phenomena-oriented re-structuring of the inventory of surface syntactic relations to enable a linguistically informed treebank transformation.

2 Background

The MTT-based annotation of the SynTagRus treebank provides various types of linguistic information. In particular, the morphological features associated with individual lexical items include the respective part of speech, and depending on it, further features like animacy, gender, number, case, degree of comparison, short form (for adjectives and participles), representation (of verbs), aspect, tense, person, and voice. In SynTagRus, sentences are represented as trees in which words are nodes and edges between them are marked with the appropriate syntactic relation. The number of nodes in the tree structure typically corresponds to the number of word tokens, and the dependencies between them are binary and oriented, i.e. linking single words rather than syntactic groups. For every syntactic group, one word (*head*) is chosen to represent it as a dependent in larger syntactic units; all other members of the group become dependents of the head word. Punctuation marks do not carry any labeling and are not included in syntactic trees.

The rich inventory of MTT surface syntactic relations – about sixty, as currently annotated in the treebank – captures fine-grained language-specific grammatical functions of the lexemes in a sentence and is traditionally divided into six major groups – i.e. actantial, attributive, quantitative, adverbial, coordinative, or auxiliary – which, in fact already provides a generic picture of abstract dependency relations and guidelines for our cross-theoretical investigation.

I. Actantial relations link a predicate word to its arguments. Prototypical instances thereof are: *predicative*, *completive, prepositional*

II. Attributive relations often link a noun to a modifier expressed by an adjective, another noun or a participle clause. Prototypical instances thereof are: *attributive, modificational, relative*

III. Quantitative relations link a noun to a quantifier or numeral, or two such words together. A prototypical instances thereof is: *quantitative*

IV. Adverbial relations link a predicate word to various adverbial modifiers. Prototypical instances thereof are: *circumstantial, parenthetic*

V. Coordinative relations serve phrases and clauses coordinated by conjunctions. Prototypical instances thereof are: *coordinative*, *coordinative-conjunctive*

VI. Auxiliary relations typically link two elements that form a single syntactic unit (e.g. an analytical verb form). Prototypical instances thereof are: *auxiliary*, *analytical*

As SynTagRus authors point out, the language-specific inventory of surface syntactic relations is not closed, as the process of data acquisition brings up rare syntactic constructions not covered by traditional grammars, which requires new syntactic link types to be introduced for make the respective syntactic structure unambiguous. Let us consider an example of the original SynTagRus annotation.

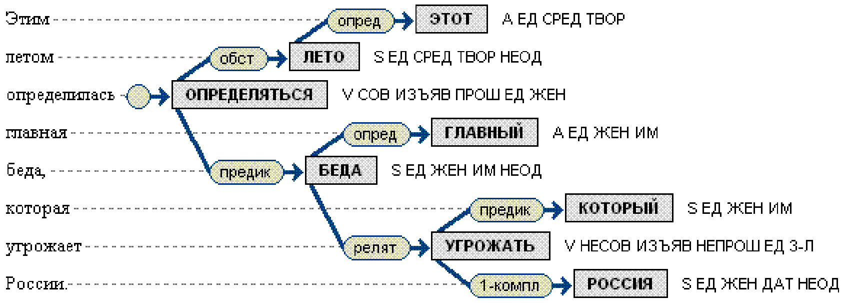

Figure 1: Original SynTagRus annotation

The sentence in Figure 1 may be indicatively translated as: “This summer took shape the main adversity that threatens Russia.” The matrix verb определилась (took shape) is in a *predicative* (предик) dependency with its subject беда (distress) and in a *circumstantial* (обст) dependency with the temporal adverbial летом (summer). The former is in a *modificational* (опред) dependency with the attributive adjective главная (main) and in a *relative* (релят) dependency with the verb of the relative clause угрожает (threatens). The latter, on the other hand, is in a modificational (опред) dependency with the demonstrative pronominal adjective этим (this). The embedded verb, in turn, is in a predicative (предик) dependency with the relative pronoun которая (which) and in a 1-completive (1-комл) dependency with its object России (Russia).

3 Treebank conversion

The conversion of the SynTagRus dependency treebank to the HPSG derivations is achieved in the following three steps. First, the dependency trees are converted into pseudo phrase structure trees by creating constituents for head words and their dependents. As the majority of the dependencies are projective, the conversion results in mostly continuous constituents. The non-continuous constituents produced from the non-projective dependencies are also preserved at this point, and will be handled in the later conversion stages. We use the Negra/Tiger XML format [11] to record the syntactic structures throughout the conversion. The format conveniently supports non-continuous constituents. The dependency relation types are also preserved in the resulting constituent tree as edge labels: the head word is governed by its upper constituent with the "HD" edge, while all its immediate dependents are governed by the upper constituent with an edge named after the corresponding dependency relation. Figure 2 shows the pseudo phrase structure tree of the example sentence from the previous section (cf. Figure 1). The constituents SP, and VP are created automatically, and named according the part of speech of the head word (i.e. "substantive" and "verb", respectively). Different bar levels of the constituents are not yet determined, and the tree structure can be rather "flat".

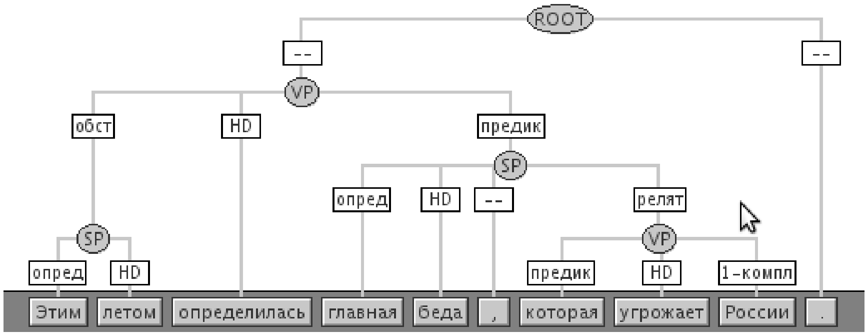

Figure 2: Converted SynTagRus format

The next step of the conversion aims to annotate the branches in the pseudo phrase structure tree with HPSG-oriented schemata. In the initial phase of the treebank conversion we work with a small set of HPSG-oriented schemata for headed phrases (cf. Table 1) which have straightforward structure-preserving correspondences in terms of MTT *surface syntactic relations*.

It is worth noting that during this conversion a language specific theory evolves. Starting from the standard HPSG inventory of schemata we eventually arrive at more fine-grained inventory modeling language specific phenomena. The resulting theory would be still HPSG inspired but also draw insight form the MTT approach.

Table 1: Initial basic inventory of HPSG phrasal schemata

<01>	HD+SBJ	*predicative*
<02>	HD+CMP	*1/2/3-completive* (with non-nominal head)*; agentive; prepositional*
<03>	HD+CMP/PRD	*copulative*
<04>	HD+CMP/ADJ	*quasi-agentive; 1-completive* (with nominal head); *elective; comparative*
<05>	HD+ADJ	*attributive, circumstantial, delimitative, relative modificational*
<06>	HD+ADJ/CPD	*compound*
<07>	HD+SPR	*quantitavie*
<08>	HD+AUX	*auxiliary*
<09>	HD+PARENTH	*parenthetical*

Schema <01> covers the *predicative* (предик) dependency holding between the verb and its subject. Schema <02> covers all *completive* (компл) dependencies of non-nominal heads as well as the *agentive* (агент) dependency introducing the "demoted" instrumental agent in passivization or nominalization constructions (i.e. equivalent to "by-phrase"), and the *prepositional* dependency between a preposition and the noun. Schema <03> covers the *copulative* (присвяз) dependency holding between a copula verb and the predicative. Schema <04> is underspecified with regard to complement or adjunct status and – with nominal heads only – covers the *completive* (компл) dependencies and the *quasi-agentive* (квазиагент) dependency to a genitive noun (i.e. equivalent to "of-phrase"), as well as the *comparative* (сравнит) dependency between a head and an indicated object of comparison and the *elective* (электив) dependency between a head and a respectively indicated set. Schema <05> covers various kinds of adjuncts corresponding to the *modificational* (опред) dependency between a noun and its agreeing (i.e. adjectival) attribute, the *attributive* (атриб) dependency between a noun and its non-agreeing (i.e. non-adjectival) attribute, the *circumstantial* (обст) dependency of a head to its adverbial modification, the *delimitative* (огранич) dependency of a particle or a quantifying adverb to the head it restricts, the *relative* (релят) dependency between the head noun and the relative clause modifying it. Schema <06> corresponds to the *compound* (композ) dependency between a head and a modifier part of a compound. Schema <07> corresponds to the *quantitative* (количест) dependency to a numeral expression. Schema <08> covers the *auxiliary* (вспом) dependency between a head and various auxiliary elements. Finally, schema <09> covers the *parenthetical* (вводн) dependency between a head and an inserted parenthetical expression which is usually divided by punctuation marks.

These schemata cover an essential part of the phenomena in the HPSG view. While some of the schemata correspond clearly with some dependency relations in a one-to-one fashion, others are not as straightforward. This reflects the asymmetry of different linguistic frameworks, and presents a major difficulty in developing conversion programs. Previous attempts in this direction usually involve the design of complex rule-based conversion heuristics (cf. [12], [9], [8]). In practice, these heuristics are also highly dependent on the annotation schema, and do not always carry linguistically interesting analyses.

In this work, we propose to use a different bootstrapping approach towards an incremental treebank conversion process. The process starts with linguists annotating instances of particular target structures, e.g. specific HPSG schemata like head-subject, head-complement, and head-adjunct. These annotations are attached to the original treebank annotation as already converted into pseudo phrase structure trees. A machine learning classifier will learn from these instances, and try to predict for the remainder of the treebank the conversion outcome. The conversion quality will be manually checked. Then the conversion results will be used as the starting point for the next (and potentially more difficult) conversion sub-step. Since for each round, we are only adding limited additional conversion decisions, annotation from a few dozen up to a few hundred instances will be enough for training the statistical classifiers.

Figure 3 shows the manual annotation of HPSG schemata on the pseudo phrase structure trees. Although the complete annotation is shown in this example, the annotators can choose to only visualize analyses they are interested in and annotate the instances they are sure about.

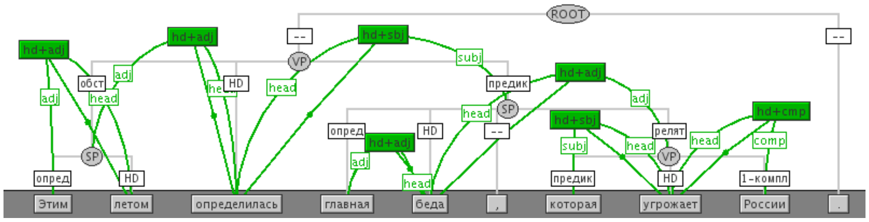

Figure 3: Manual HPSG-oriented meta-annotation

These annotations are then sent to train the statistical classifier, which is applied to disambiguate the mappings from dependency relations to the HPSG schemata. We use a maximum entropy-based classifier (TADM, http://tadm.sourceforge.net). The effective features for schemata classification include the part-of-speech of the head and daughter in the pseudo phrase structure tree, the dependency label, together with the sibling non-head daughters. The results are illustrated in Figure 4. While the edge labels now bear more resemblance to HPSG, the phrases structures are still flat.

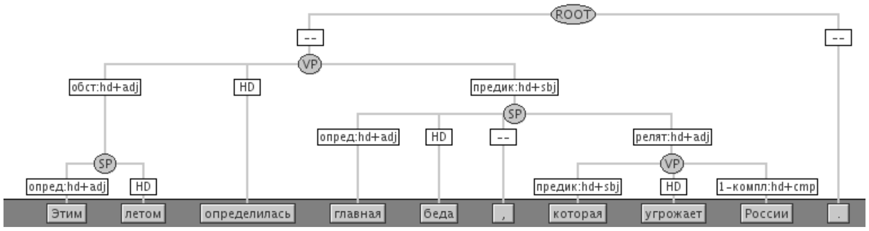

Figure 4: Automatic annotation with statistical classifier

Our experiments resulted in adequate automatic meta-annotation of the development corpus summarized in Table 2. In general, the assignment of core HPSG schemata, i.e. head-subject (with 172 occurrences), head-complement (with 354 occurrences), head adjunct (with 521 occurrences), and head-specifier (with 15 occurrences), is convincingly stable and highly conform to the initial setup of basic phrasal types outlined in Table 1. The same is true of the head-complement/adjunct schema (with 149 occurrences), which we introduced to account for the systematic functional status under-specification of a nominal head's dependents in *quasi-agentive* and *completive* surface syntactic relations.

Table 2: Experimental automatic meta-annotation results

Development corpus statistics		**Dependency**	**Schema**
242	`опред:hd+adj`	*modificational*	HD+ADJ
174	`предл:hd+cmp`	*prepositional*	HD+CMP
172	`предик:hd+sbj`	*predicative*	HD+SBJ
145	`1-компл:hd+cmp`	*1-completive*	HD+CMP
112	`обст:hd+adj`	*circumstantial*	HD+ADJ
105	`огранич:hd+adj`	*delimitative*	HD+ADJ
94	`квазиагент:hd+cmp/adj`	*quasi-agentive*	HD+CMP/ADJ
47	`1-компл:hd+cmp/adj`	*1-completive*	HD+CMP/ADJ
38	`атриб:hd+adj`	*attributive*	HD+ADJ
28	`2-компл:hd+cmp`	*2-completive*	HD+CMP
15	`количест:hd+spr`	*quantitative*	HD+SPR
15	`релят:hd+adj`	*relative*	HD+ADJ
13	`вводн:hd+parenth`	*parenthetic*	HD+PARENTH
8	`2-компл:hd+cmp/adj`	*2-completive*	HD+CMP/ADJ
8	`сравнит:hd+adj`	*comparative*	HD+ADJ
6	`3-компл:hd+cmp`	*3-completive*	HD+CMP
6	`присвяз:hd+cmp/prd`	*copulative*	HD+CMP/PRD
3	`вспом:hd+aux`	*auxiliary*	HD+AUX
2	`композ:hd+adj/cpd`	*compound*	HD+ADJ/CPD
1	`агент:hd+cmp`	*agentive*	HD+CMP
1	`электив:hd+cmp/adj`	*elective*	HD+ADJ

The assignment of other schemata, i.e. head-parenthetical (with 13 occurrences), head-predicative-complement (with 6 occurrences), head-auxiliary (with 3 occurrences), and head-adjunct-in-compound (with 2 occurrences), appears to give quite satisfactory results too. The *delimitative* (огранич) dependency, which involves heterogeneous non-head categories, has received in the experimental results an interpretation mainly as a head-adjunct structure (105 occurrences). Nevertheless, the theoretical question arises of whether to re-interpret this surface syntactic relation – at least in cases involving quantifying particles (negative, interrogative, topicalising, etc.) – as a head-marker structure. Also, a linguistically motivated interpretation of both *comparative* (сравнит) and *elective* (электив) surface syntactic relations would favor under-specification of the non-head component with regard to its complement or adjunct status, which corresponds to the head-complement/adjunct schema.

There are, in fact, a whole bunch of surface syntactic relations that have been intentionally excluded from the current experiment and, hence, got no meta-annotation in terms of HPSG schemata – cf. Table 3. For examples of individual dependency types refer to [7]. These are all, to a various degree, non-trivial cases, with the most representative group being the treatment of coordination phenomena.

Table 3: Dependencies currently excluded from meta-annotation

Development corpus statistics		**Dependency**
98	сочин	*coordinative*
90	соч-союзн	*conjunctive-coordinative*
30	подч-союзн	*conjunctive-subordinative*
26	сент-соч	*sentential-coordinative*
15	разъяснит	*expository*
7	аппоз	*appositive*
7	эксплет	*expletive*
5	сравн-союзн	*conjunctive-comparative*
4	примыкат	*adjunctive*
3	1-несобст-компл	*1-nonintrinsic-competive*
3	4-компл	*4-completive*
3	аналит	*analytical*
3	инф-союзн	*conjunctive-invinitival*
3	кратн	*multiple*
3	распред	*distributive*
2	длительн	*durative*
2	оп-опред	*descriptive-modificational*
2	пролепт	*proleptic*
2	соотнос	*correlational*
1	2-несобст-компл	*2-nonintrinsic-completive*
1	компл-аппоз	*completive-appositive*
1	ном-аппоз	*nominative-appositive*
1	об-аппоз	*detached-appositive*

Inasmuch as coordination relations are not dependencies in the strict sense of the word, their handling is always one way or another conventionalized in dependency grammar approaches. In SynTagRus, according to the Meaning Text Theory, the first coordination member is the head and is attached to the common parent, i.e. to the word governing the entire coordination. Each other coordination member (including conjunctions) is attached to the previous one, with the edges between coordination members being labeled with the *coordinative* (`сочин`) or the *conjunctive-coordinative* (`соч-союзн`) dependencies. With respect to common dependents, i.e. words depending on all coordination members, one particular solution has been favored in SynTagRus, namely, that these are attached to the nearest coordination member, often to the first one, with the other coordination members, including conjunctions, being attached to the respectively preceding one. The systematic source of ambiguity – whether a dependent of a coordination member actually refers to the whole coordination or only to that one member – is thus deliberately avoided in SynTagRus.

All the HPSG schemata we have are binary structures. This is because they are always more informative than flat structures involving more than two daughters. Also binary structure bears more resemblance to the dependency relations between pairs of words. For this reason, we need to further binarize the pseudo phrase structure trees. This turns out to be a non-trivial step for languages with relatively free word order. As there is less constraints over the linear precedence between constituents, it is hard to hard-wire schema priorities directly. Similar to the previous step, we start by annotating some of the binarization preferences by hand, and hope that the regularities will be then transferred to the remainder of the corpus. For example, in Figure 5, the left-most binarization annotation indicates that the verbal head will pick up the right-adjacent subject before combing with the modifying noun phrase to its left.

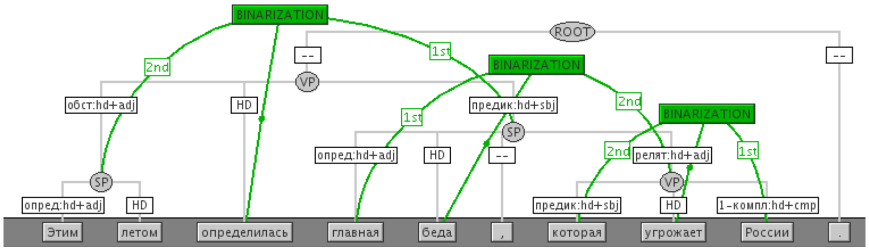

Figure 5: Manual binarization

The learning of such regularities turns out to be more difficult too. For a constituent with a head `H` together with additional `m` pre-head daughters and `n` post-head daughters, there are in total `(m+n)!/(m! n!)` possible binarizations of the tree. While a simple classifier is employed to guess the structure, better formulation of this as a machine learning task will be investigated in the future. Figure 6 shows an example of the binarization result.

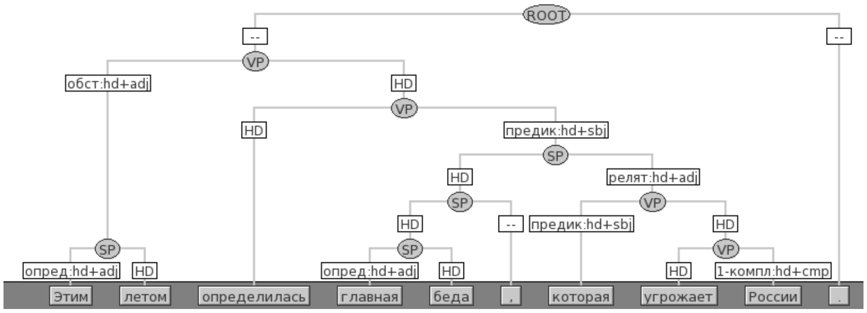

Figure 6: Final structure

It is worth pointing out that the resulting derivation trees only reflect partial view on a complete HPSG analysis. In our case, both corpus and grammar are under parallel development, and draw insights from each other's progress. In future development, we will apply the constraints of the HPSG schemata in the hand-written grammar to the derivation trees. The HPSG signs will be instantiated through this process, allowing us to acquire detailed lexicon for our grammar. For the core grammar development, we are using the DELPH-IN grammar engineering platform (http://www.delph-in.net/), which supports the dynamic evolution of both grammar and treebank as in the LinGO Redwoods approach [14].

4 Conclusion

In our view, phenomena-oriented re-structuring of the inventory of surface syntactic relations has the potential of enabling linguistically informed treebank transformation. In this contribution we've presented the first results of creating a constituency treebank of Russian by converting the detailed dependency annotation of SynTagRus to schematic HPSG derivations, taking into account the genuine hierarchy of surface syntactic relations.

The general setup is sketched in Figure 7. We have no access to the grammar and the lexicon of the ETAP-3 linguistic processor [1]. Nevertheless we can utilize the structured linguistic knowledge contained in it, working directly with the output of the system as provided in the syntactic annotation of the SynTagRus treebank. The resulting converted treebank, which we tentatively call SynTagRus++, is of crucial importance for the implementation of a broad-coverage precision Russian resource grammar in the context of creation of open-source Slavic grammatical resources [5]. The latter initiative aims at ensuring an optimal and efficient grammar engineering cycle through dynamic coupling of treebanks, computer grammars and other relevant resources for the Slavic language family.

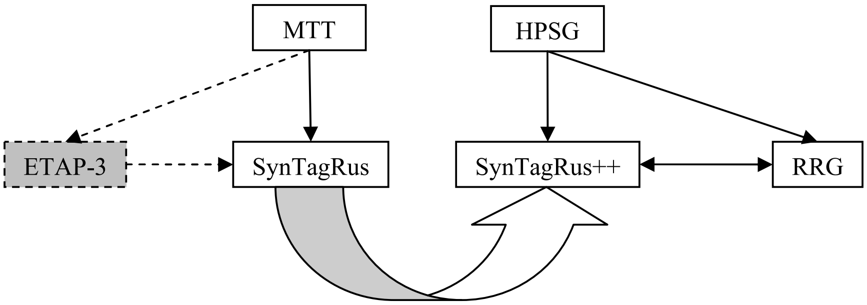

Figure 7: General setup

On the theoretical level our work contributes towards a conceptual alignment between two established linguistic theories: MTT and HPSG. This is a novel and extremely challenging topic, which calls for treebank-supported in-depth cross-theoretical investigations.

References

[1] Apresian, Juri, Igor Boguslavsky, Leonid Iomdin, Aleksandr. Lazursky, Viktor Sannikov, Viktor Sizov, and Leonid Tsinman. (2003) ETAP-3 linguistic processor: A full-fledged NLP implementation of the MTT. In *First International Conference on Meaning-Text Theory*. p. 279–288.

[2] Apresjan, Juri, Igor Boguslavsky, Boris Iomdin, Leonid Iomdin, Andrei Sannikov, and Victor Sizov (2006) A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects. In *The fifth international conference on Language Resources and Evaluation, LREC 2006*. Genoa, Italy.

[3] Avgustinova, Tania and Yi Zhang (2009) Developing a Russian HPSG based on the Russian National Corpus. In *DELPH-IN Summit*. Barcelona.

[4] Avgustinova, Tania and Yi Zhang (2009) Exploiting the Russian National Corpus in the Development of a Russian Resource Grammar. In *Proceedings of the RANLP-2009 Workshop on Adaptation of Language Resources and Technology to New Domains*. Borovets, Bulgaria.

[5] Avgustinova, Tania and Yi Zhang (2009) Parallel Grammar Engineering for Slavic Languages. In *Workshop on Grammar Engineering Across Frameworks at the ACL/IJCNLP*. Singapore.

[6] Boguslavsky, Igor, Svetlana Grigorjeva, Nikolai Grigorjev, Leonid Kreidlin, and Nadezhda Frid (2000) Dependency treebank for Russian: Concept, tools, types of information. In *COLING*. p. 987-991.

[7] Boguslavsky, Igor, Ivan Chardin, Svetlana Grigorjeva, Nikolai Grigoriev, Leonid Iomdin, Leonid Kreidlin, and Nadezhda Frid. (2002) Development of a dependency treebank for Russian and its possible applications in NLP. In *Third International Conference on Language Resources and Evaluation (LREC-2002)*. Las Palmas. p. 852–856.

[8] Cahill, Aoife, Mairéad McCarthy, Josef van Genabith, and Andy Way (2002) Automatic Annotation of the Penn-Treebank with LFG F-Structure Information. In *LREC 2002 workshop on Linguistic Knowledge Acquisition and Representation - Bootstrapping Annotated Language Data*. Third International Conference on Language Resources and Evaluation, post-conference workshop: ELRA – European Language Resources Association. p. 8-15.

[9] Hockenmaier, Julia and Mark Steedman (2007) CCGbank: a corpus of CCG derivations and dependency structures extracted from the Penn Treebank. Computational Linguistics 33(3): p. 355-396

[10] Mel'cuk, Igor' A. (1988) Dependency Syntax: Theory and Practice. State University of New York Press.

[11] Mengel, Andreas and Wolfgang Lezius (2000) An XML-based encoding format for syntactically annotated corpora. In *Proceedings of the Second International Conference on Language Resources and Engineering (LREC 2000)*. Athens. p. 121-126.

[12] Miyao, Yusuke, Takashi Ninomiya, and Junichi Tsujii (2005) Corpus-oriented Grammar Development for Acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank, In *Natural Language Processing - IJCNLP 2004, LNAI3248, Hainan Island, China*, Keh-Yih Su, Jun'ichi Tsujii, Jong-Hyeok Lee, and Oi Yee Kwong, Editors. Springer-Verlag. p. 684-693.

[13] Nivre, Joakim, Igor Boguslavsky, and Leonid Iomdin (2008) Parsing the SynTagRus Treebank. In *COLING*. p. 641–648.

[14] Oepen, Stephan, Dan Flickinger, Kristina Toutanova, and Christopher D. Manning (2004) LinGO Redwoods. A rich and dynamic treebank for HPSG. Journal of Research on Language and Computation. 2(4): p. 575-596

[15] Tesnière, Lucien (1959) Éléments de syntaxe structurale. Paris Klincksieck.