Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2023"

June 14–16, 2023

The problem of linguistic markup conversion: the transformation of the Compreno markup into the UD format

Alexandra Ivoylova
RSUH
Moscow, Russia
a.m.ivoylova@gmail.com

Darya Dyachkova
RSUH
Moscow, Russia
d.dyachkova@bk.ru

Maria Petrova
A4 Technology
Moscow, Russia
g-fox-ive@mail.ru

Mariia Michurina
RSUH
Moscow, Russia
marimitchurina@gmail.com

Abstract

The linguistic markup is an important NLP task. Currently, there are several popular formats of the markup (Universal Dependencies, Prague Dependencies, and so on), which are mostly focused on morphology and syntax. Full semantic markup can be found in the ABBYY Compreno model. However, the structure of the format differs significantly from the models mentioned above. In the given work, we convert the Compreno markup into the UD format, which is rather popular among NLP researchers, and enrich it with the semantical pattern.

Compreno and UD present morphology and syntax differently as far as tokenization, POS-tagging, ellipsis, coordination, and some other things are concerned, which makes the conversion of one format into another more complicated. Nevertheless, the conversion allowed us to create the UD-markup containing not only morpho-syntactic information but also the semantic one.

Keywords: Compreno, semantic markup, Universal Dependencies
DOI: 10.28995/2075-7182-2023-22-191-199

Проблемы конвертации лингвистической разметки: конвертация формата Compreno в UD-формат

Ивойлова А.М.
РГГУ
Москва, Россия
a.m.ivoylova@gmail.com

Дьячкова Д.С.
РГГУ
Москва, Россия
d.dyachkova@bk.ru

Петрова М.А.
A4 Technology
Москва, Россия
g-fox-ive@mail.ru

Мичурина М.А.
РГГУ
Москва, Россия
marimitchurina@gmail.com

Аннотация

Лингвистическая разметка является актуальной задачей NLP. В настоящее время существует несколько популярных форматов подобной разметки (Universal Dependencies, Prague Dependencies и др.), при этом в фокусе их внимания находятся, в первую очередь, морфология и синтаксис. Одним из немногих форматов, предлагающих не только морфо-синтаксическую, но и семантическую разметку, является формат ABBYY Compreno, однако в структурном отношении данный формат существенно отличается от указанных выше моделей. В предлагаемой работе делается попытка представить разметку Compreno в более привычном для пользователей формате UD и дополнить данный формат семантической разметкой.

Представление морфологии и синтаксиса в UD и Compreno имеет ряд значимых различий, касающихся, в числе прочего, токенизации, POS-tagging, эллипсиса, сочинения и других явлений, что создает определенные сложности при конвертации. Тем не менее, конвертация Compreno в UD позволила получить полную многоуровневую UD-разметку, содержащую как морфо-синтаксическую, так и семантическую информацию.

Ключевые слова: Compreno, семантическая разметка, Universal Dependencies

1 Introduction

Morphological, syntactic and semantic labelling is an essential part of natural language processing pipeline. A need for the universal multilanguage markup format has been acknowledged for a long time; one of the most known projects of creating such a format is the Universal Dependencies (UD) project (De Marneffe et al., 2006), although UD encompasses morphosyntax only.

As for the semantics, there is no markup standard so far which would be widely acknowledged. Currently, several projects deal with semantic labels, and some of them are meant for integral three-level labelling, for instance, Prague Dependencies (Hajic et al., 2001), or the ETAP system (Boguslavsky, 1999). Nevertheless, none of these projects provide both laconic and integral labelling format.

An attractive model in this respect seems the ABBYY Compreno model (Anisimovich et al., 2012; Petrova, 2014) which is capable to perform a full-scale morphosyntactic and semantic labelling. Its advantage is the ability to provide a complete well-structured semantic markup, which includes not only arguments, but also adjuncts, modifiers, and other dependencies. Besides, it has special means of handling non-tree links, such as ellipsis or dislocation. However, Compreno has its own drawbacks: first, the semantic part of the markup is too detailed which makes the markup too complicated; second, the formal structure of the markup format has some peculiarities.

Our primary goal is thereby to develop a new labelling standard that would benefit both from the conciseness of UD and the thoroughness of Compreno system. To achieve it, we decided to adopt the UD format for the morphosyntactic markup part and to enrich it with the simplified Compreno semantic markup. This task, in turn, demanded the conversion of the Compreno markup format into UD.

The elaboration of the integral markup standard and, especially, the semantic markup standard is a part of the Compreno-Based Linguistic Data (CoBaLD) Annotation Project which includes the creation of a fully-labelled Russian dataset [1] of approximately 400,000 tokens as well, containing news texts from the (now defunct) NewsRu.Com site. For more information on the standard and the dataset[2], see (Petrova et al., 2023).

At the first stage, the corpus was automatically annotated by the Compreno parser and checked manually by professional linguists.

At the second stage, the morphosyntactic part of the markup was automatically converted into the UD format. The conversion was partly checked as well. To evaluate the quality of the conversion, we checked abound 10% of the dataset. The percent of labels modified by different groups of annotators in manually checked automatic conversion varies from 5 to 10%, which means that the total quality of the conversion is close to 95%.

After the conversion, the UD markup was supplemented with the semantic pattern - word meanings for each token and the semantic relations between the constituents.

In the current paper, we focus on one important part of the work - the conversion of one format into another and the challenges we encountered solving this problem.

2 Related Work

The need to have a standardized format of natural text labelling (at first, POS-tagging) appeared when the first corpora were created. The pioneers of POS-tagging are the creators of the Brown corpus, the Lancaster/Oslo-Bergen corpus, the University of Pennsylvania corpus (UPenn) and others. The first language for labelling was English. A comprehensive table of rival English POS-tags can be found in (Atwell, 2008).

However, it turned out difficult to use the English POS-tags for other languages, so the attempts were made to create language-specific tagsets, such as (Bar-Haim et al., 2008) for Modern Hebrew or (Diab, 2007) for Arabic. Approximately at the same time, the UD project started. Its creators strove to develop a universal standard which could be applied to any language and which would combine both morphological and syntactic features.

[1] https://github.com/compreno-semantics/compreno-corpus

[2] The access to the dataset is provided according to the CC BY-NC 4.0 License which allows non-commercial use.

On the other side, semantic labelling formats were being developed as well, starting with the well-known Universal Networking Language (UNL) (Uchida and Zhu, 2001), and onto more recent projects like Universal Decompositional Semantics (UDS) (White et al., 2016) and Universal Conceptual Cognitive Annotation (UCCA) (Abend and Rappoport, 2013). The semantic projects, however, concentrated on the semantics itself. Some of them do not involve morphosyntactic parsing at all (the already mentioned UCCA and Abstract Meaning Representations (Banarescu et al., 2013) are the examples).

As for the Russian language, the only two projects aimed at the semantic parsing are the ETAP project and the above-mentioned Compreno project. For the moment, there are only a few Russian datasets labelled in UD (e.g., Taiga (Shavrina and Shapovalova, 2017) or SynTagRus (Boguslavsky, 2014; Droganova and Zeman, 2016; Droganova et al., 2018); as for the semantic labelling, no datasets are available.

3 Overview of the Format

Our markup format is derived from the UD format [3] and represents a format very similar to the well-known CONLL which is a syntactic parse tree with semantic data included. The main representation principles in UD are the following: a sentence is separated by a newline, any table row contains ten columns, and the columns include token ID, form, lemma, universal POS-tags and language-specific POS-tags, grammatical features, dependency head and dependency relation.

Unlike UD, the Compreno format represents sentences in a tree-like structure (see fig. 1):

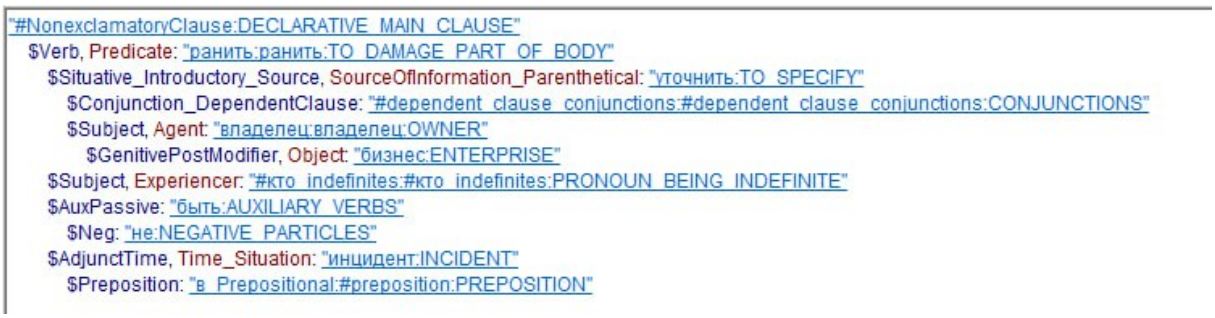


Figure 1: Compreno format: tree structure for *Как уточнил владелец бизнеса, никто не был ранен в инциденте.* 'As the business owner clarified, no one was injured in the incident.'

For our standard, we adopted the UD table format, but replaced the last two (usually empty, especially in the Russian UD corpora) columns with semantic slots and semantic classes taken from the Compreno format.

The Compreno model presents words in the form of a thesaurus-like semantic tree, which consists of universal semantic classes - semantic fields, filled with lexical contents in each language incorporated in the model. The total number of the classes is more than 200 000. For the current work, we used the simplified version of the hierarchy, cut to hyperonym classes only (about 1000 classes). For details, see (Petrova et al., 2023) and the relevant fragment of the hierarchy on Github[4].

Semantic slots, in turn, correspond to semantic roles, which define the semantic relations between the core and the dependent elements, including actants such as Agent or Experiencer, characteristics, adjuncts (time, condition, concession, etc.), and so on. Unlike syntactic roles, semantic ones can have different syntactic realizations, for instance, all bracketed constituents in "I will come [tomorrow]", "I will come [after sunset]", and "I will come [when the clock strikes twelve]" correspond to Time slot. Or, subject-Agent in active voice and by-Agent in passive voice correspond to one Agent slot. The list of the slots can be found on Github[5].

An example of the labelled text can be seen in fig. 2.

[3]The complete information on the UD tagset which was implemented here may be found at https://universaldependencies.org/).

[4]https://github.com/compreno-semantics

[5]https://github.com/compreno-semantics/compreno-corpus/blob/main/semantic_slots.xlsx

```
# text = Как уточнил владелец бизнеса , никто не был ранен в инциденте.
1   Как как SCONJ   _   _   2   mark    _   CONJUNCTIONS
2   уточнил уточнить    VERB    _   Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Act  9   parataxis   Parenthetical   VERBAL_COMMUNICATION
3   владелец    владелец    NOUN    _   Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing   2   nsubj   Agent   HUMAN
4   бизнеса бизнес  NOUN    _   Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing   3   nmod    Object_Situation    BUSINESS
5   ,   ,   PUNCT   _   _   2   punct   _   _
6   никто   никто   PRON    _   Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing|Polarity=Neg  9   nsubj:pass  Experiencer BEING
7   не  не  PART    _   Polarity=Neg    8   advmod  _   PARTICLES
8   был быть    AUX _   Aspect=Imp|Gender=Masc|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Act    9   aux:pass    _   AUXILIARY_VERBS
9   ранен   ранить VERB    _   Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part|Voice=Pass 0   root    Predicate   TO_DAMAGE_PART_OF_BODY
10  в   в   ADP _   _   11  case    _   PREPOSITION
11  инциденте   инцидент    NOUN    _   Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing   9   obl Time    FACT_INCIDENT
12  .   .   PUNCT   _   _   9   punct   _   _
```

Figure 2: Our format: markup for *Как уточнил владелец бизнеса, никто не был ранен в инциденте.* 'As the business owner clarified, no one was injured in the incident.'

4 Compreno2UD Converter

The annotation of the dataset in such a format demanded the creation of the automatical converter which would transform Compreno morphosyntantic markup into UD.

The conversion program consists of several blocks. These blocks include original markup extraction, syntax and morphology conversion. The semantic layer is simply added over the resulting markup as it does not have to be converted.

The conversion pipeline is as follows:

- Labelled and manually checked texts are extracted from the Compreno system with the help of an API. On this stage, we get separate semantic and morphosyntactic data (morphological and syntactic labels are not divided technically);
- The extracted data is parsed and handed onto the syntactic module;
- Both the results of the syntax conversion and the original morphological data are passed to the morphological module, where tokenization and lemmatization issues are solved as well, and the results of both stages are merged;
- The semantic markup is merged with the results of the conversion.

The conversion of morphology and syntax can be performed in any order, so the reason for the syntax being converted first is purely technical.

Now let us consider the syntax and the morphology conversion in more detail, especially as far as the asymmetry between Compreno and UD is concerned.

5 Syntax

The description of the syntactic parsing in Compreno can be found in (Anisimovich et al., 2012). Shortly, the parser builds the dependency tree for each sentence, where each node is provided with the necessary grammatical features (both morphological and syntactic). Each dependency is marked with the surface slot (or syntactic role) such as Subject, Object_Direct, Object_Instrumental, and so on.

Compreno restores all elided nodes (such as copulas, for instance) and has a special set of labels for dislocation cases.

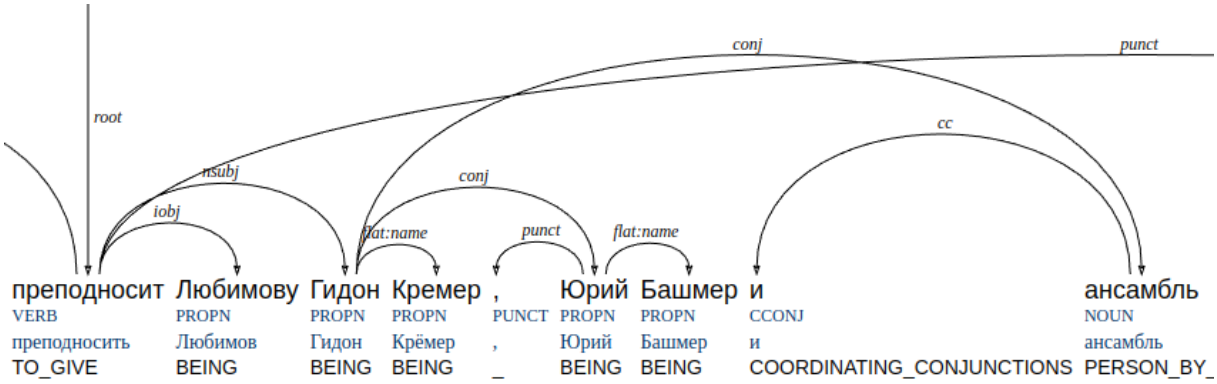
The process of syntax conversion is divided into two parts: the conversion of the heads and the conversion of the relations. Technically, the conversion of the heads must be done first, as the information about the heads is used during the relations conversion.

5.1 Dependency Heads

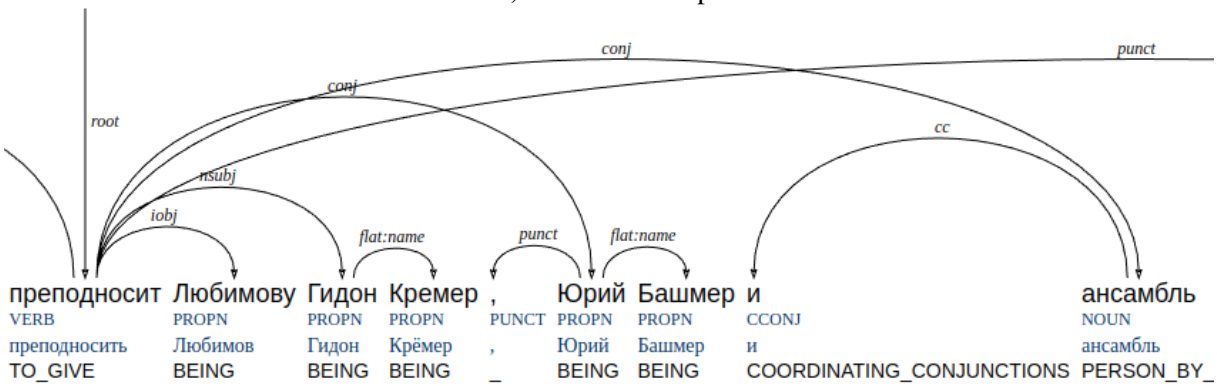
The conversion of the dependency heads from the Compreno format into the UD one seems quite straightforward to a large extent. However, there are several asymmetry cases dealing with ellipsis, coordination and movements. Moreover, the punctuation marks in Compreno are not regarded as separate nodes unlike it is in UD. Therefore, the main issues for the conversion were as follows:

- Punctuation marks had to get their dependency heads with the help of the rule-based algorithm which took into account the heads of the tokens on both sides of the mark in question;

- The cases of the elided heads were treated as close to the UD format as possible, with several rules according to the UD documentation on ellipsis. Nevertheless, this part of the conversion is prone to errors due to its rule-based nature;
- The copula in Compreno is the head of its clause, that is, in the sentence *Вася был студентом* 'Vasya was a student' the root is *был* 'was'. In UD, the root is the complement of the copula (*студентом* 'student');
- The preposition *согласно* 'according to' in Compreno is considered the head; in UD, it behaves like any other preposition, being dependent on its noun. Oddly enough, it is the only case of this kind we found in our data;
- The coordination is treated differently in UD and in Compreno - this divergence can be seen in fig. 3: as one can see, in UD, the coordinated elements depend on the first element of the coordination (*Гидон*), while in Compreno, all coordinated elements depend on the similar core (here - the verb *преподносит*). We adopted the UD concept.



a) The UD concept



b) The Compreno concept

Figure 3: Conjuncts representation (in UD style): a) the UD concept; b) the Compreno concept

For each case, we scripted the conversion to be as close to UD as possible.

Some of these differences were easy to eliminate, while the others involved a lot of discussions.

5.2 Dependency Relations

The dependency relations in UD cannot be treated as purely syntactic, as they often consider some semantic features. Compreno, on the other hand, has a strict distinction between syntax and semantics.

Therefore, the conversion of the Compreno format into UD included creating the set of rules which take into account various syntactic features, data on dependency heads and sometimes morphological and semantic categories. In most cases, it was not hard to align Compreno categories with UD dependency relation types.

The most difficult types for conversion appeared to be the following.

First of all, there are *obj*, *iobj* and *obl* relations in UD, and the distinctions between them are not purely syntactic: if there are two or more objects, one should choose the *obj* relation for the closest object and the *iobj* relation for the rest; the 'closeness' of the objects is hard to determine automatically.

Secondly, we could not truly define the *dislocated* tag, as there are no consistent features for it in Compreno, or they are difficult to derive.

We also did not implement the conversion rules for the *list*, *goeswith* and *reparandum* tags, as there were none in our dataset (typos in the data were corrected during manual semantic labelling).

6 Morphology

The morphology level is represented by POS-tags and grammatical features. As simple as it may seem, the attempts to develop a POS tagset for any language inevitably reveal some dubious areas. The same turned out to be true for the approach to the grammatical features. Both - the sets of POS and the sets of grammatical features do not coincide in the given formats. In general, we tried to follow the UD guidelines in most cases in order to be as consistent as possible with the format.

6.1 POS-tagging

Key differences between Compreno and UD in this respect are the following:

- There is no *Predicative* as a POS-tag in UD, but there is one in Compreno. This tag is assigned in cases like *Мне нужно идти* 'I must go'. Following the UD principles, we chose to convert the predicatives to adverbs, though it may be an arguable decision;
- There is no *Determiner* tag in Compreno, but there is one in UD. Taking into account that there is a closed set of tokens marked as determiners in UD (words like *мой, свой, этот, какой* 'mine', 'own', 'this', 'which'), we converted them using a list of tokens and a syntactic rule 'the head of the token in question must be a noun' (the rule helps to avoid placing the Determiner-tag in cases like *это было вчера* 'this happened yesterday');
- There is no POS-tag for proper nouns in Compreno, while there is one in UD. However, there are special grammatical features (grammemes) for proper nouns such as *Proper*, so UD POS-tags are set according to them;
- There are also inconsistencies with ordinal numerals (*одиннадцатый* 'eleventh'), which are tagged as numerals in Compreno, and as adjectives in UD (we convert them as adjectives);
- Some tokens in the Compreno format get a special POS-tag 'Invariable', which does not correspond to any of the UD tags; usually, these are discourse units and parenthetical constructions, for instance, *впрочем* 'however'. We created a special list of such tokens in order to process them according to UD.

6.2 Grammatical Features

We also encountered some asymmetry cases while mapping grammatical features. The information encoded in UD by one tag is sometimes distributed between several tags in Compreno, for example:

	UD	Compreno
Short forms	Variant=Short	ParticipleShortForm AdjectiveShortForm
Abbreviations	Abbr=Yes	Abbreviation Lex_Abbreviation Lex_KgSm Lex_LetterAbbreviation Lex_LetterDotAbbreviation

Furthermore, some grammatical categories are divided into morphological and syntactic ones in Compreno: for example, there are *Gender* and *SyntacticGender* tags for the grammatical category of gender. If some token has the *Gender=Common* tag, it means that its gender can change based on the context, that is, that the same token can function both as Feminine and Masculine. As a rule, these are surnames: *Шеварднадзе, Кириленко*, but the words like *убийца, камикадзе* also fall here, as well as some

foreign names: *Associated Press, УЕФА*. In this case, we use the information from the *SyntacticGender* tag.

Another example is the pronoun *себя* 'oneself' and alike. It does not have number and gender tags in most markups, but it gets them in Compreno in accordance with the semantic component as it inherits these categories from the controller of *себя*. In our resulting format, we decided to keep only information about case, as it is done in other UD-corpora.

6.3 Tokenization and Lemmatization

The conversion task also included the processing of lemmas and tokens since the principles of tokenization and in a lesser extent lemmatization are different in UD and in Compreno.

One of the prominent lemmatization differences is that the Compreno system puts verb lemmas in the perfect aspect, while in UD, verb lemma should have the same aspect as its form in a sentence. In order to comply with the UD format, we created a list of all verbs with both variants and restored their correct lemmas based on the aspect tag. This may be an arguable decision, as there are discussions on whether one should consider verb aspect an inflectional feature or not.

6.4 Re-tokenization

Technically, the most difficult part of the job was re-tokenizing sentences as tokenization rules for UD and Compreno differ significantly. For instance, the UD format implies that there cannot be a space inside a token, while Compreno treats many idiomatic and syntactically opaque expressions, such as кроме того 'besides', при этом 'moreover', and so on as a single unit.

To cope with this asymmetry in the conversion process, it was necessary not only to divide or split tokens in accordance with the UD standard, but also to decide what tags the parts of the split tokens should inherit. To divide and merge the tokens, we used the dictionary which was partly based on the list in the SynTagRus to UD conversion repository (257 tokens) and subsequently changed it and supplemented with new cases (now 389 tokens). Further, the list will be filled with all the tokens which include a space in the Compreno database. In this dictionary the head of a bigram is defined as a new head of a split token, and only this head inherits the semantic class label, while the others get none.

The splitting of foreign words and tokens like '*1990-1991*', company names, and time intervals demanded creating some rules as well. Such cases are innumerable and cannot be taken into account in any list, so we split them rule-based.

Another re-tokenization task was the merge of cases which could not be processed with the help of a list. The following token groups were merged:

- immutable parts of compound words (*авиа, фильмо, шведско*);
- model or product names (*Ту-34, Ил-76*);
- numerals with endings (*70-й, 19-го*).

Re-tokenization also invokes an issue of distributing semantic categories on the last stage of the markup conversion. When a token is split, its semantic slot and class are assigned to its part which would be the syntactic head in the split construction; the rest of the parts would get blanks. For example, the token *кроме того* 'besides' would be split into two parts, where *того* is the head of *кроме*. The semantic class DISCOURSIVE_UNITS would be assigned to *того*, as it is the class of the whole token *кроме того* in Compreno.

7 Compreno vs UD: challenges

As the conversion showed, there are two problems to work on further in more detail, both concern non-tree syntax.

First, UD does not restore the syntactic zeros, which leads to 'unnatural' dependencies. For example, in the phrase *Спортсменка, показав второй результат на первом участке, вылетела с трассы на втором* 'The athlete, having shown the second result in the first section, flew off the track in the second' the token *втором* 'second' depends on the verb and is marked with the *obl* relation, substituting its elided head. In the Compreno model, the elided head *участке* 'section' would be restored here, and

every constituent would get its correct tags. This difference was really hard to take into account during the conversion, so there must be inconsistencies in our dataset with such cases. As a consequence, the *orphan* relation was implemented only partially.

Second is the conversion of the constituents dislocation. For instance, let us take the sentence *Как подсказывает опыт, в классические шахматы лучшую игру демонстрируют сильнейшие шахматисты* 'As experience suggests, in classical chess the best play is demonstrated by the strongest chess players'. The correct head for the constituent *в шахматы* 'in chess' would be the head *игра* 'play' (and it is so in the Compreno format). Such information can be taken from the semantic structure of the sentence built by the parser, however, the current version of the converter does not process this information properly, and the head is assigned wrongly as *демонстрируют*. This problem is going to be solved by re-working the conversion script.

8 Further Developments

As a natural development of our work, we consider modifying our current markup format by adding the elided heads. This task will be probably tricky, as there is no satisfying concept for the labelling of the elided heads for now: it is difficult to include such nodes in the current CONLL format, because they do not have phonetically expressed forms.

As for the architecture of the converter, we will improve its work with the original parsed trees from Compreno in order to restore ellipsis and to label the dependency heads correctly in case of dislocation. Needless to say, we will focus on the correction of any bugs found in the current version of the converter.

9 Conclusion

It is commonly known that automatic conversion of any type of linguistic markup is a difficult task. In the current paper, we have shown the conversion process of the Compreno markup format into the UD morphosyntactic markup standard. The full description of the automatic conversion blocks - from tokenization to syntax - has been provided, with the focus on some fundamental differences and inconsistencies between the standards. The result of our work is a fully-labelled dataset for the Russian language which includes approximately 400,000 tokens. The dataset markup follows UD guidelines in the morphosyntactic part and is supplemented with the semantic pattern. Further work presupposes modifications in the syntax level such as restoring ellipsis.

References

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (ucca). // *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 228–238.

KV Anisimovich, K Yu Druzhkin, KA Zuev, FR Minlos, MA Petrova, and VP Selegei. 2012. Syntactic and semantic parser based on abbyy compreno linguistic technologies. // *Proc Dialogue, Russian International Conference on Computational Linguistics*, P 91–103.

ES Atwell. 2008. Development of tag sets for part-of-speech tagging.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. // *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, P 178–186.

Roy Bar-Haim, Khalil Sima'An, and Yoad Winter. 2008. Part-of-speech tagging of modern hebrew text. *Natural Language Engineering*, 14(2):223–251.

Igor Boguslavsky. 1999. Translation to and from russian: the etap system. // *EAMT Workshop: EU and the new languages*.

Igor Boguslavsky. 2014. Syntagrus–a deeply annotated corpus of russian. *Les émotions dans le discours-Emotions in Discourse*, P 367–380.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. // *Lrec*, volume 6, P 449–454.

Mona Diab. 2007. Improved arabic base phrase chunking with a new enriched pos tag set. // *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, P 89–96.

Kira Droganova and Daniel Zeman. 2016. Conversion of syntagrus (the russian dependency treebank) to universal dependencies. Technical report, Technical report, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data conversion and consistency of monolingual corpora: Russian ud treebanks. // *Proceedings of the 17th international workshop on treebanks and linguistic theories (tlt 2018)*, number 155, P 53–66. Linköping University Electronic Press Linköping, Sweden.

Jan Hajic, Barbora Vidová-Hladká, and Petr Pajas. 2001. The prague dependency treebank: Annotation structure and support. // *Proceedings of the IRCS Workshop on Linguistic Databases*, P 105–114.

M Petrova, A Ivoylova, I Bayuk, D Dyachkova, and M Michurina. 2023. The CoBaLD Project: the creation and application of the full morpho-syntactic and semantic markup standard. // *International Conference on Computational Linguistics and Intellectual Technologies «Dialog»*.

MA Petrova. 2014. The compreno semantic model: the universality problem. *International Journal of Lexicography*, 27(2):105–129.

Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning:“taiga” syntax tree corpus and parser. // *Proceedings of “CORPORA-2017” International Conference*, P 78–84.

Hiroshi Uchida and Meiying Zhu. 2001. The universal networking language beyond machine translation. // *International Symposium on Language in Cyberspace, Seoul*, P 26–27.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. // *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, P 1713–1723.