ST PETERSBURG STATE UNIVERSITY

PROCEEDINGS

OF THE INTERNATIONAL CONFERENCE

«CORPUS LINGUISTICS–2021»

July 1–3, 2021, St. Petersburg

СКИФИЯ
принт

St. Petersburg
2021

Организационный комитет конференции
«Корпусная лингвистика–2021»

В. П. Захаров (председатель), И. В. Азарова,
Е. Л. Алексеева, Л. Н. Беляева, А. О. Гребенников,
О. Н. Камшилова, О. В. Митренина, О. А. Митрофанова,
И. С. Николаев (зам. председателя), В. И. Фирсанова,
М. В. Хохлова, А. В. Чижик

Программный комитет конференции
«Корпусная лингвистика–2021»

В. П. Захаров (председатель), Н. Абдурахмонова (Узбекистан), И. В. Азарова,
Е. Л. Алексеева, В. А. Баранов, Л. Н. Беляева, В. Бенко (Словакия),
О. В. Блинова, С. Ю. Богданова, Н. В. Борисов, В. В. Бочаров,
Р. фон Вальденфельс (Германия), А. М. Галиева, Р. Гарабик (Словакия),
А. Горак (Чехия), А. О. Гребенников, Д. О. Добровольский,
Т. Елинек (Чехия), Л. Л. Иомдин, Е. Каллас (Эстония), О. Н. Камшилова,
К. И. Коваленко, М. С. Коган, А. В. Колмогорова, Е. Н. Колпачкова,
М. В. Копотев (Финляндия), М. Кршен (Чехия), Д. А. Кочаров,
А. М. Лаврентьев (Франция), У. Лоу (Великобритания), О. Н. Ляшевская,
О. В. Митренина, О. А. Митрофанова, М. М. Михайлов (Финляндия),
А. Д. Москвина, О. А. Невзорова, И. С. Николаев (зам. председателя),
В. Нозеда (Италия), Х. Нэси (Великобритания), К. Пала (Чехия),
В. Петкевич (Чехия) (зам. председателя), А. Ч. Пиперски,
Л. В. Рычкова (Беларусь), С. О. Савчук, В. П. Селегей, Д. В. Сичинава,
О. Скривнер (США), В. Д. Соловьев, А. Стефанович (Германия),
Е. В. Суворина, Ю. Тао (Китай), Н. Тюренн (Франция), М. В. Хохлова,
А. Я. Шайкевич, С. А. Шаров (Великобритания), Т. Ю. Шерстинова,
С. Эйден (Франция), Е. В. Ягунова, М. Якубичек (Чехия)

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ТРУДЫ
МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ
«КОРПУСНАЯ ЛИНГВИСТИКА–2021»

1–3 июля 2021 г., Санкт-Петербург

СКИФИЯ
принт

Санкт-Петербург
2021

ББК 81.1
Т78

Ответственный редактор издания
В. П. Захаров

Т78 **Труды международной конференции «Корпусная лингвистика–2021».** — СПб.: Издательство Скифия-принт, 2021. — 396 с.

ISSN 2412-9623
ISBN 978-5-98620-557-1

Сборник содержит материалы докладов, представленных на научной конференции «Корпусная лингвистика-2021» 1–3 июля 2021 г. в Санкт-Петербурге.

Создание и использование корпусов текстов является одним из приоритетных направлений в современной лингвистике. Проведение конференции по данной тематике знакомит ученых с современными разработками и новыми технологическими решениями в этой области, а также способствует обобщению опыта научных исследований по корпусной лингвистике.

ББК 81.1

ISSN 2412-9623
ISBN 978-5-98620-557-1

© Авторы, 2021

ОГЛАВЛЕНИЕ

ПЛЕНАРНЫЕ ДОКЛАДЫ/KEYNOTE TALKS

D. Zeman

ENHANCED UNIVERSAL DEPENDENCIES: THE CURRENT STATE AND OUTLOOK 9

С. О. Савчук

НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА В ЗЕРКАЛЕ СТАТИСТИКИ 18

С. П. Тимошенко, Л. Л. Иомдин, С. А. Гладилин, Е. С. Иншакова

СИНТАГРУС В СОСТАВЕ НКРЯ: НОВЫЕ ВОЗМОЖНОСТИ 31

В. И. Беликов, А. Д. Верещагина, В. П. Селегей

НА ГРАНИЦАХ ЛЕКСИКОНА: ДИФФЕРЕНЦИАЛЬНЫЕ КОРПУСНЫЕ ИССЛЕДОВАНИЯ ПАРЕМИЙНОГО ФОНДА РЯ 44

СЕКЦИОННЫЕ ДОКЛАДЫ/Section TALKS

L. Abalo-Dieste, J. Pérez-Guerra

PASSIVISATION AND RELATIVISATION AS COLLOQUIALISATION STRATEGIES IN PRESENT-DAY ENGLISH: A CORPUS-BASED STUDY 56

L. N. Beliaeva, O. N. Kamshilova

MT RESULTS AND PARALLEL SCIENTIFIC TEXT CORPORA FOR LEXICOGRAPHY 65

V. Bénet, M. Silberztein

CORPUS PROCESSING: THE LINGUISTIC APPROACH DEVELOPING RESSOURCES FOR RUSSIAN, APPLICATIONS TO LINGUISTICS AND LANGUAGE TEACHING 74

I. Chiari, M. Bader, A. Salem, L. Squillante

USING CORPORA IN BUILDING A MULTILINGUAL GLOSSARY OF MIGRATION 84

N. Cortegoso Vissio, V. Zakharov

A RULE-STOCHASTIC HYBRID POS-TAGGER FOR SRANAN TONGO WITH MINIMAL LEXICON AND TRAINING DATASET 95

A. R. Gatiatullin., D. S. Suleymanov, N. A. Prokopyev, M. M. Saifullin

"TURKIC MORPHEME" PORTAL AS A TOOL FOR UNIFICATION OF ANNOTATION SYSTEM FOR TURKIC ELECTRONIC CORPORA 104

M. A. Klimova, V. K. Smilga, D. A. Overnikova

USING AN ERROR-ANNOTATED LEARNER CORPUS (REALEC) IN DDL LESSONS 112

A. Orenha-Ottaiano, M. E. O. de Oliveira Silva

A CORPUS-BASED PLATFORM OF MULTILINGUAL COLLOCATIONS DICTIONARIES (PLATCOL): SOME LEXICOGRAPHICAL ASPECTS AIMING AT PRE- AND IN-SERVICE TEACHERS 122

N. Perkova, D. Sitchinava

ON THE SWEDISH-RUSSIAN PARALLEL CORPUS AND ITS POSSIBLE APPLICATIONS (WITH THE FOCUS ON SEVERAL SWEDISH CONSTRUCTIONS) 133

A. Ch. Piperski

SIMPLICITY BEATS SOPHISTICATION: AN EVALUATION OF ADJUSTED FREQUENCY MEASURES 140

Н. Л. Аванесян, А. М. Чеповский, Т. Ю. Шерстинова, Ф. Н. Соловьев, Д. Ю. Чуйкин

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ ЧАСТОТНЫХ СЛОВАРЕЙ ЛИНГВИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ РУССКОЙ ПРОЗЫ 1900–1930 ГГ. В ДИНАМИКЕ 149

Л. Д. Бадмаева

ПРОБЛЕМЫ МОРФОЛОГИЧЕСКОГО ОПИСАНИЯ ЛЕКСЕМ СТАРОМОНГОЛЬСКИХ ТЕКСТОВ 158

В. А. Баранов

ПАРАЛЛЕЛЬНЫЙ КОРПУС СЛАВЯНСКИХ СПИСКОВ ПАРИМЕЙНИКА: МАТЕРИАЛ И ПОСТАНОВКА ЗАДАЧИ 167

О. В. Блинова, Н. А. Тарасов

СЛОЖНОСТЬ РУССКИХ ПРАВОВЫХ ТЕКСТОВ: МЕТОДЫ ОЦЕНКИ И ЯЗЫКОВЫЕ ДАННЫЕ 175

Н. В. Богданова-Бегларян

МАРКЕРЫ-КСЕНОПОКАЗАТЕЛИ В РУССКОЙ ПОВСЕДНЕВНОЙ РЕЧИ: АННОТИРОВАНИЕ РЕЧЕВОГО КОРПУСА, ТИПОЛОГИЯ И КОЛИЧЕСТВЕННЫЕ ДАННЫЕ 183

Е. А. Вольф, Ю. О. Короткова, К. И. Семенов

АВТОМАТИЧЕСКАЯ РАЗМЕТКА ЗАИМСТВОВАНИЙ ИЗ РУССКОГО ЯЗЫКА В КИТАЙСКИХ ТЕКСТАХ: ПРОБЛЕМЫ СЛОВОДЕЛЕНИЯ И МОРФОПАРСИНГА 191

А. О. Гребенников, Т. Г. Скребцова

КОРПУС РУССКИХ РАССКАЗОВ (1900–1930). УСТОЙЧИВОСТЬ ЛИНГВОСТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК 201

Д. О. Добровольский, Анна А. Зализняк

ПАРАЛЛЕЛЬНЫЙ КОРПУС КАК ИНСТРУМЕНТ СЕМАНТИЧЕСКОГО АНАЛИЗА: НЕМЕЦКИЙ МОДАЛЬНЫЙ ГЛАГОЛ *SOLLEN* *209*

О. В. Донина

АВТОМАТИЗАЦИЯ ВЫГРУЗКИ РЕЗУЛЬТАТОВ ПОИСКА В КОРПУСАХ ЧЕТВЕРТОГО ПОКОЛЕНИЯ 219

А. В. Захарова

ВОЗМОЖНОСТЬ КОРПУСА: ОТПРАВНЫЕ ТОЧКИ И ОПЫТ СОЗДАНИЯ КОРПУСОВ МУЛЬТИМОДАЛЬНЫХ ТЕКСТОВ 225

С. С. Земичева, А. А. Васильченко

ДРУГ, ПОДРУГА, ТОВАРИЩ, ЗНАКОМЫЙ: СПЕЦИФИКА СЕМАНТИКИ И ФУНКЦИОНИРОВАНИЯ В ДИАЛЕКТНОЙ РЕЧИ (ПО КОРПУСНЫМ ДАННЫМ) 232

В. И. Зубов, Е. И. Риехакайнен

ВМЕСТЕ ИЛИ ВРОЗЬ: НЕОДНОСЛОВНЫЕ ЕДИНИЦЫ В КОРПУСАХ И В МЕНТАЛЬНОМ ЛЕКСИКОНЕ НОСИТЕЛЯ РУССКОГО ЯЗЫКА 240

А. Е. Колесников, Л. А. Малахов

ДИГИТИЗАЦИЯ ПЕЧАТНЫХ МОЛДАВСКИХ ДИАЛЕКТНЫХ ЗАПИСЕЙ 248

М. В. Копотев, О. В. Кисселев, А. А. Климов

CAT&KITTENS: КОРПУС РУССКИХ АКАДЕМИЧЕСКИХ ТЕКСТОВ
И ОСНОВАННЫЕ НА НЕМ ИНСТРУМЕНТЫ АНАЛИЗА СТУДЕНЧЕСКИХ РАБОТ 255

Н. А. Коротаев

ПОИСК В МУЛЬТИКАНАЛЬНОМ КОРПУСЕ: СОДЕРЖАТЕЛЬНЫЕ ЗАДАЧИ
И ТЕХНИЧЕСКАЯ РЕАЛИЗАЦИЯ 263

И. Л. Корецкая

ПРОБЛЕМЫ ЯПОНСКОЙ КОРПУСНОЙ ЛИНГВИСТИКИ 272

Е. А. Корсакова

СОСТАВЛЕНИЕ КОРПУСА НАУЧНЫХ ПУБЛИКАЦИЙ В СФЕРЕ ОПТИКИ 281

У. Е. Кочеткова, П. А. Скрелин

ПАРАЛИНГВИСТИЧЕСКИЕ ЯВЛЕНИЯ ПРИ ВЫРАЖЕНИИ ИРОНИИ В РУССКОМ ЯЗЫКЕ
(НА МАТЕРИАЛЕ МУЛЬТИМЕДИЙНОГО КОРПУСА ИРОНИЧЕСКИХ ВЫСКАЗЫВАНИЙ) 288

Г. И. Кустова

ИНФИНИТИВНЫЕ КОНСТРУКЦИИ С ПРЕДИКАТИВАМИ РАЗНЫХ СЕМАНТИЧЕСКИХ КЛАССОВ
В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА 296

А. М. Лаврентьев, Л. А. Курышева

ЛИНГВИСТИЧЕСКАЯ ОБРАБОТКА ЦИФРОВЫХ ИЗДАНИЙ РУССКИХ ТЕКСТОВ XVIII ВЕКА 306

А. Н. Лапошина, Т. С. Веселовская, Л. Ю. Жильцова, О. Ф. Купрещенко, М. Ю. Лебедева

КОРПУСНОЕ УЧЕБНИКОВЕДЕНИЕ: В ПОИСКАХ ОБЪЕКТИВНЫХ КРИТЕРИЕВ ОЦЕНКИ УРОВНЯ
УЧЕБНИКОВ ДЛЯ БИЛИНГВОВ 313

С. А. Мелещева, П. М. Эйсмонт

ЭЛЕМЕНТЫ НЕВЕРБАЛЬНОГО ПОВЕДЕНИЯ В ПРОЦЕССЕ ПОРОЖДЕНИЯ
НЕПОДГОТОВЛЕННОГО НАРРАТИВА РУССКОЯЗЫЧНЫМИ ДЕТЬМИ
(КОРПУСНОЕ ИССЛЕДОВАНИЕ) 322

И. С. Пименов

СПЕЦИФИКА АРГУМЕНТАЦИОННОГО АННОТИРОВАНИЯ НАУЧНЫХ
И НАУЧНО-ПОПУЛЯРНЫХ ТЕКСТОВ 330

В. И. Подлесская

АППРОКСИМАТОР *ЧТО НАЗЫВАЕТСЯ*: ЭТАПЫ ПРАГМАТИЗАЦИИ В ЗЕРКАЛЕ НКРЯ 338

Г. Саяма

ВЛИЯНИЕ ЧАСТОТНОСТИ НА ФОРМЫ ПАДЕЖНЫХ ОКОНЧАНИЙ *-АМ, -АМИ, -АХ*
(*-ЯМ, -ЯМИ, -ЯХ*) 348

Ю. Тао, В. П. Захаров

КОРПУСНЫЙ АНАЛИЗ АНТРОПОМОРФНЫХ МЕТАФОР В КИТАЙСКОМ И РУССКОМ
ПОЛИТИЧЕСКОМ ДИСКУРСЕ ПО ТЕМЕ «ОДИН ПОЯС — ОДИН ПУТЬ» 355

М. Ю. Товкес

ИССЛЕДОВАНИЕ ГЕНДЕРНЫХ СТЕРЕОТИПОВ ПОЛИТИЧЕСКОГО ДИСКУРСА С ПРИМЕНЕНИЕМ КОРПУСНЫХ ТЕХНОЛОГИЙ 362

М. В. Хохлова, И. Д. Мамаев

РАЗРАБОТКА БАЗЫ ДАННЫХ КОЛЛОКАЦИЙ: ОБЗОР ЗОЛОТОГО СТАНДАРТА НА ПРИМЕРЕ АТРИБУТИВНЫХ СЛОВОСОЧЕТАНИЙ 370

А. Н. Чевелева, Э. С. Клышинский

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ КОНСТРУКЦИЙ ДЛЯ ПОВЕРХНОСТНОГО СИНТАКСИЧЕСКОГО АНАЛИЗА 380

П. Чжан

КОРПУСНЫЙ АНАЛИЗ ЛЕКСИЧЕСКОГО НАПОЛНЕНИЯ КОНЦЕПТА «ГОСУДАРСТВО» В РУССКОЙ И КИТАЙСКОЙ ЯЗЫКОВЫХ КАРТИНАХ МИРА 387

ПЛЕНАРНЫЕ ДОКЛАДЫ

KEYNOTE TALKS

D. Zeman

ENHANCED UNIVERSAL DEPENDENCIES: THE CURRENT STATE AND OUTLOOK[1]

Abstract. Universal Dependencies (UD) is a multilingual collection of corpora featuring morphological and syntactic annotation in a unified style. We discuss an optional layer of deep-syntactic annotation in UD, called Enhanced Universal Dependencies. We survey the existing enhanced representation as of release 2.8 and consider two possible future expansions: semi-automatic addition of existing enhancement types to new languages, and addition of new enhancement types.
Keywords. Dependency syntax, deep syntax, multilingual corpora, gapping, coordination, coreference.

1. Introduction

Universal Dependencies[2] (UD) [Nivre et al. 2020; de Marneffe et al. 2021] is an international community project that strives to define a unified morpho-syntactic annotation scheme applicable to all natural languages, and to collect corpora (treebanks) annotated following that scheme. It started with the first version of annotation guidelines in 2014, and with the first release of 10 treebanks in January 2015; after six years, release 2.8 of UD boasts about 202 treebanks for 114 languages from 24 different families. Some treebanks are just tiny samples of less than thousand tokens while others contain over a million tokens; the total size of the collection amounts to 27 million.

UD has become an indispensable resource for research on multi-lingual natural language processing, especially morphological tagging and syntactic parsing. It has been also used in many linguistic studies, in particular in linguistic typology. Two large CoNLL shared tasks were organized in 2017 and 2018 to evaluate parsing systems on UD data [Zeman et al. 2018].

[1] This work was supported by the Grant No. GX20-16819X (LUSyD) of the Czech Science Foundation (GAČR).

[2] https://universaldependencies.org/

The morphological annotation in UD includes the lemma, universal part-of-speech tag (UPOS), morphological feature-value pairs, and possibly another tag from a treebank-specific tag set (XPOS). The UPOS tag must be picked from a fixed set of 17 categories; any finer distinctions, if desirable, are encoded in the morphological features (in addition to universally defined features, treebanks may also use language-specific values and features). All UD treebanks must have at least the UPOS tags manually checked. Lemmas, features and XPOS are optional and in a few UD corpora they have been assigned automatically.

The *basic* syntactic representation is a rooted dependency tree where every word/node (except the root) has one parent node. Each relation in the tree is labeled with its type; while the main types come from a fixed set of 37 universal relations, it is possible to define language-specific subtypes. Like UPOS tags, the parent nodes and the relation types are manually checked in all UD treebanks.

Examples of UD basic trees with UPOS tags are given in Figures 1 (English) and 2 (Russian), respectively. The sentences are parallel and so is their syntactic annotation: Relations between content words are identical in both structures, although English also has a number of additional relations be-

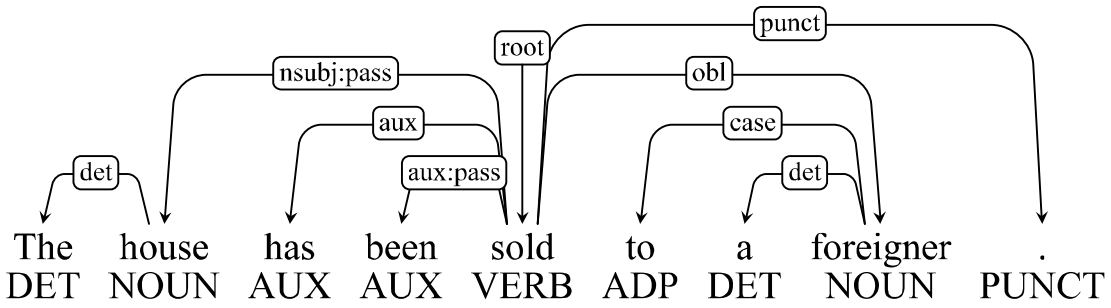

Fig. 1. **Basic UD tree of the English sentence *The house has been sold to a foreigner***

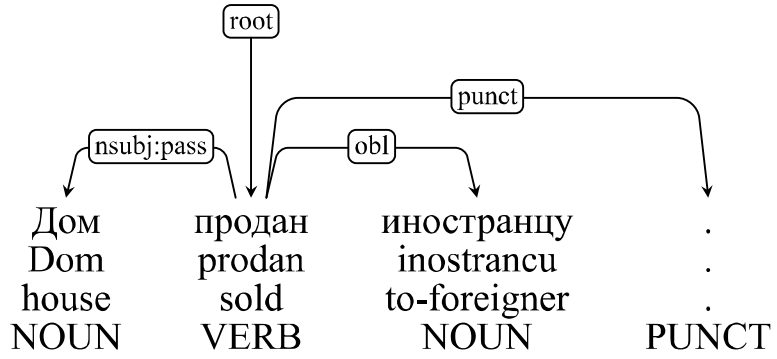

Fig. 2. **Basic UD tree of the Russian sentence *Дом продан иностранцу* (The house has been sold to a foreigner)**

tween content and function words, which are not present in Russian. Such parallelism is possible thanks to the fact that content words are attached higher in UD, and function words (articles, prepositions, auxiliaries etc.) are normally attached as leaf nodes — an approach that is relatively uncommon in dependency frameworks outside UD.

2. Enhanced Representation

Besides the basic syntactic representation, UD may optionally contain an *enhanced* dependency structure, which is still a directed and rooted graph, but not necessarily a tree. The enhanced UD layer was proposed by [Schuster and Manning 2016] but its first official specification appeared in the version 2 of the UD guidelines [Nivre et al. 2020: § 3.4]. The purpose of the enhanced representation is to facilitate downstream language understanding tasks by making certain relations explicitly annotated. In most cases, the enhanced graph is only a moderate modification of the basic tree. There are six types of enhancements defined in the guidelines (a UD treebank may annotate only some types and ignore the others):

1) gapping (empty nodes for elided predicates);
2) parent of coordination (propagated relations to non-first conjuncts);
3) shared dependent of coordination (propagated relations from non-first conjuncts);
4) external subject of a controlled or raised verb;
5) relative clause (modified noun attached instead of the relative pronoun, thus forming a directed cycle);
6) case information in the relation label.

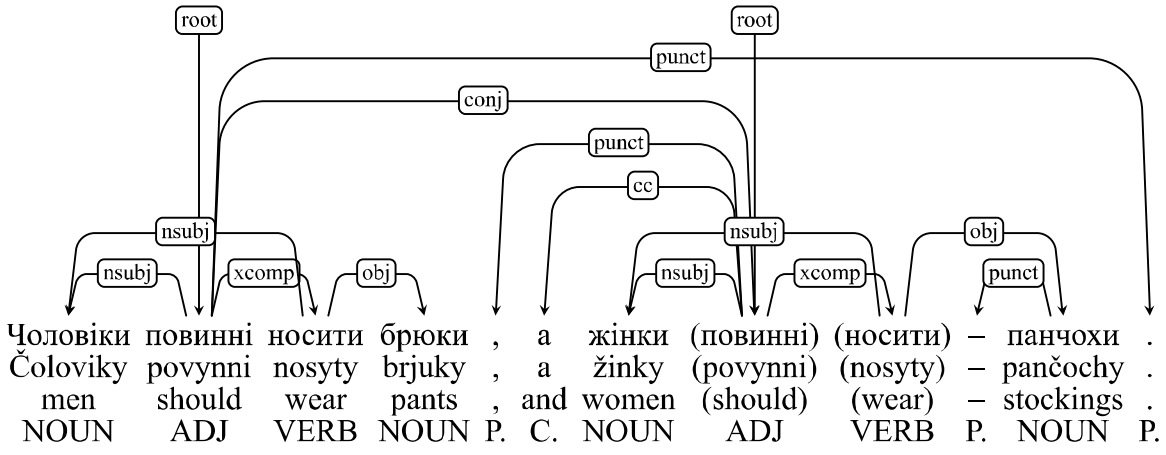

Fig. 3. **Enhanced UD graph of the Ukrainian sentence *Чоловіки повинні носити брюки, а жінки — панчохи* (Men should wear pants and women should wear stockings)**

Figure 3 illustrates three enhancement types. There are two 'empty' nodes that represent elided predicates in a gapping construction, *повинні* "should" and *носити* "wear". Propagation of coordination parent results in the second root relation, pointing to the empty node representing the second instance of *повинні*. And finally, both *чоловіки* "men" and *жінки* "women" have two incoming relations each, making them subject not only of *повинні* but also of *носити*. Figure 4 illustrates enhanced representation of relative clauses; note the directed cycle between *szamponu* "shampoo" and *myje* "washes". For more details on the six enhancement types, see [Droganova and Zeman 2019; Nivre et al. 2020].

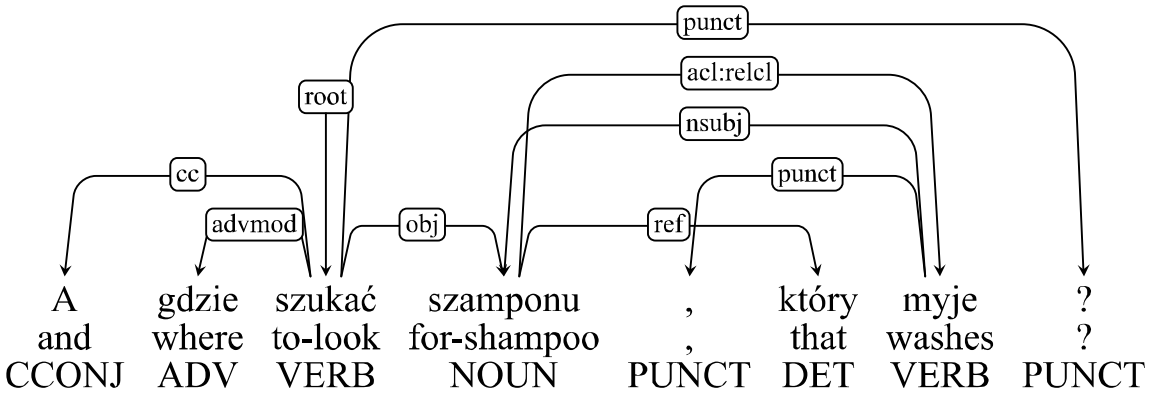

Fig. 4. **Enhanced UD graph of the Polish sentence *A gdzie szukać szamponu, który myje?* (And where to look for shampoo that works?)**

Some of the enhancements can be computed deterministically from the basic tree, as they really just foreground information that is already there. This is the case with coordination parents and case-enhanced relation labels. Other types will benefit from manual disambiguation but they still can be approximated algorithmically using relatively simple heuristics. At the other end of the scale, distinguishing between shared dependents of coordination and private dependents of the first conjunct clearly requires extra knowledge and may be sometimes hard even for a human annotator. Several tools have been proposed that use heuristics to compute some of the enhancements from the basic tree; see [Nivre et al. 2018] for a comparison. Among the UD treebanks that currently have some enhanced dependency structures, little annotation (if any at all)[3] has been added manually. In some cases, the treebanks were converted from non-UD annotation schemes where the extra information was available. In others, heuristic enhancers

[3] Exact information about the origin of enhanced annotation in individual treebanks is not available.

were employed. Table 1 shows statistics of enhancement types in UD 2.8, which is the most recent release at the time of writing. Overall, 30 treebanks of 18 languages have at least one enhancement type, and 15 treebanks of 8 languages have all six types.

Table 1. **Overview of enhancements in UD 2.8. Number of nodes includes empty nodes. Enhanced relations per 1,000 nodes: G = gapping relations (i.e., to or from empty nodes); P = relations propagated from parents of coordination; S = relations propagated to shared dependents of coordination; X = relations to controlled external subjects; R = relations added in relative clauses (usually 2 for each clause); C = relation labels enhanced with case information**

Treebank	Nodes	G	P	S	X	R	C
Arabic PADT	282,460	2	48	10		16	286
Belarusian H.	305,406	4	44	0	6	16	148
Bulgarian BT.	156,149		7	4	3	16	118
Chukchi HSE	6,207	15					1
Czech CAC	495,497	15	65	22	5	24	184
Czech FicTree	167,371	11	55	19	5	21	127
Czech PDT	1,509,052	11	45	11	7	23	173
Czech PUD	18,623	4	39	1	7	26	172
Dutch Alpino	208,747	2	14	5	6	18	130
Dutch LassyS.	98,242	5	25	8	2	12	138
English EWT	254,857	1	21	9	14	11	128
English GUM	134,553	2	26	8	13	11	143
English GU.R.	16,286		19	11	18	14	123
English PUD	21,183	1	21	7	12	16	150
Estonian EDT	438,175	4	0	0		15	42
Estonian EWT	68,968	6				14	53
Finnish PUD	15,817	1					
Finnish TDT	202,453	6	33	32	5	0	
Italian ISDT	298,380	1	24	5	5	20	169
Latvian LVTB	252,961	13	36	32	9		158

End of the Table 1

Treebank	Nodes	G	P	S	X	R	C
Lithuanian A.	70,051	0	77	27	6	11	233
Polish LFG	130,967		12	8	7	0	85
Polish PDB	350,036		41	20			
Polish PUD	18,389		37	20			
Russian Syn.	1,107,741	4					3
Slovak SNK	106,184	5	39	7	4	15	151
Swedish PUD	19,085	2	22	8	11	28	158
Swedish Tal.	96,859	2	34	9	11	22	149
Tamil TTB	9,581		25			7	271
Ukrainian IU	122,324	10	51	8	9	17	

Similarly to the basic dependencies, there are parsing models that can generate the enhanced graphs for previously unseen text. Some successful parsers take advantage of the fact that many enhancements can be guessed based on the basic tree and combine a tree parsing model with enhancing heuristics. Enhanced UD parsers have been evaluated in two shared tasks run in connection with the IWPT 2020 and 2021 conference [Bouma et al. 2020].

3. What Is Next?

As Table 1 clearly shows, the enhanced representation, being optional, is only available for a fraction of the UD treebanks, and it does not grow as quickly as the data with the basic representation. [Droganova and Zeman 2018] note that this is unlikely to change, as more complex annotation requires more annotation effort, and it is thus difficult to get sufficient manpower to annotate data in a new language. They propose to at least apply a heuristic enhancer to all UD treebanks after each release and make this data available to the users. In addition, they propose heuristics to normalize syntactic alternations such as passive vs. active diathesis; they call the resulting data Deep UD, to distinguish it from the Enhanced UD defined in the official UD guidelines. Various other "enhanced-plus" variants have been proposed by [Schuster and Manning 2016] and others, but they are

not (yet?) approved as a part of UD. Even the existing guidelines sometimes leave room for multiple interpretations, leading to enhancement 'sub-types' that only appear in some treebanks.

The treatment of relative clauses can be extended to attributively used participles, as in the French example in Figure 5, where *fusée* "rocket" is modified by the participle *pouvant* "able", and at the same time it is also annotated as the external subject of the participle (as well as its complement infinitive *menacer* "threaten").

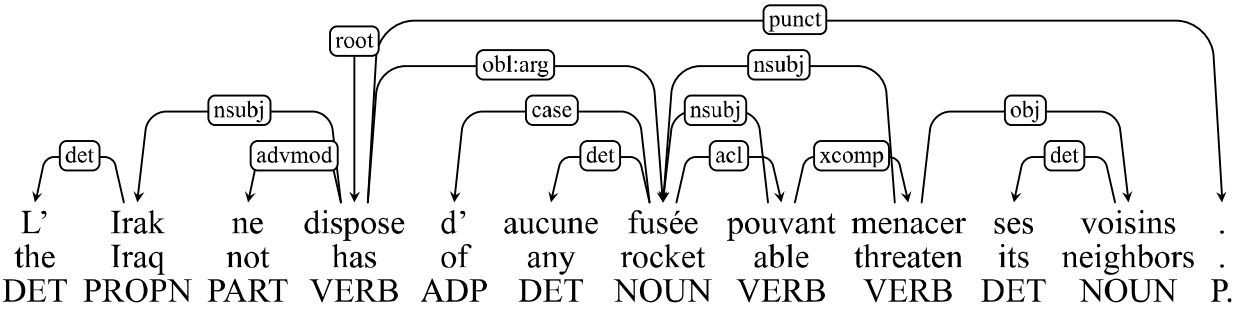

Fig. 5. **Enhanced UD graph of the French sentence *L'Irak ne dispose d'aucune fusée pouvant menacer ses voisins* (Iraq has no rockets that could threaten its neighbors)**

In pro-drop languages, empty nodes could help restore the coreference of controlled subjects if the main subject is missing, as the empty node (*on*) is used in Figure 6.

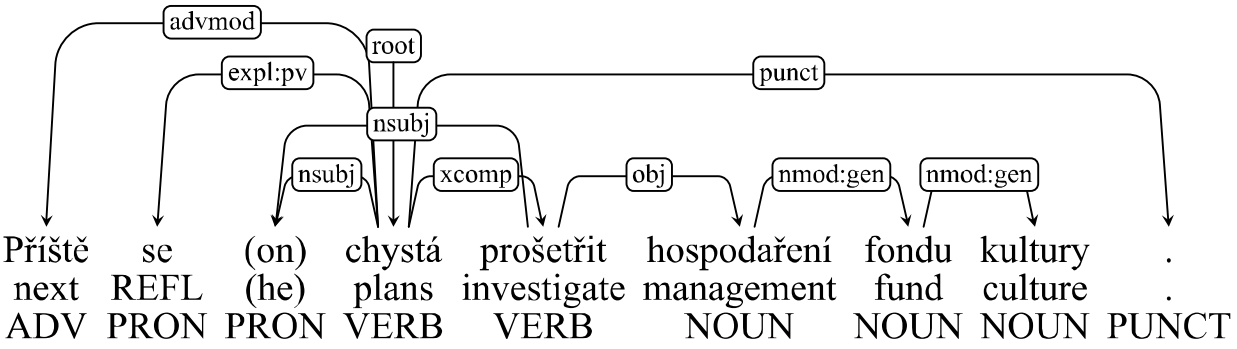

Fig. 6. **Enhanced UD graph of the Czech sentence *Příště se chystá prošetřit hospodaření fondu kultury* (Next time they are going to investigate the management of the culture fund)**

Finally, empty nodes in enhanced UD have been used to show the attachment of constituents that are incorporated in the verb and lack a node in the basic representation, like the Chukchi adverb *ныткы* "again" in Figure 7.

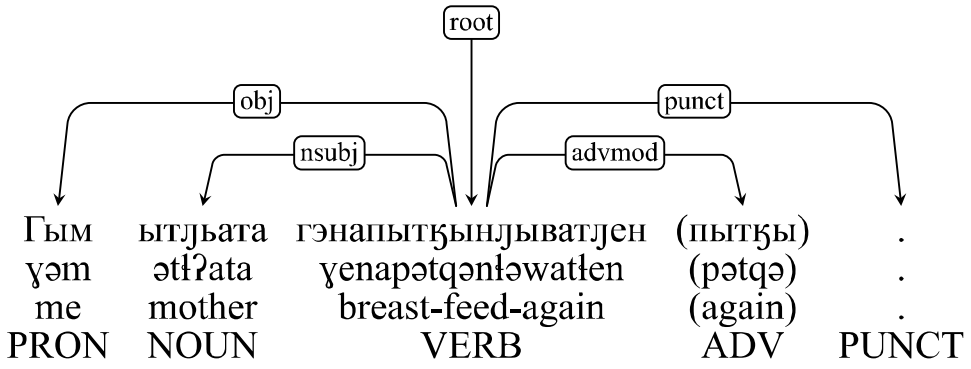

Fig. 7. **Enhanced UD graph of the Chukchi sentence *Гым ытԓьата гэнапытҕынԓываттен* (My mother was breast-feeding me again)**

4. Conclusion

We have presented the current state of the *enhanced representation* in Universal Dependencies. Being an optional and more complex annotation layer, it is only available for a fraction of the UD treebanks. Fortunately, a significant part of it can be computed or estimated with simple heuristics from the basic representation.

The guidelines for enhanced graphs are not considered as frozen as the basic guidelines in the UD community, and some details are still being elaborated as more languages are added. There is room for future additions of new variants of existing enhancement types or even completely new types with the same general motivation: to make otherwise implicit syntactico-semantic relations explicit and thus more easily accessible for language understanding applications.

References

1. *Bouma G., Seddah D., Zeman D.* (2020), Overview of the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies. In: Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies, pp. 151–161, ACL, Stroudsburg, PA, USA. URL: https://www.aclweb.org/anthology/2020.iwpt-1.16/ (date of access: 01.07.2021).
2. *de Marneffe M.-C., Manning C., Nivre J., Zeman D.* (2021), Universal Dependencies. In: Computational Linguistics. doi.org/10.1162/coli_a_00402
3. *Droganova K., Zeman D.* (2019), Towards Deep Universal Dependencies. In: Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019), Paris, France, pp. 144–152, URL: https://www.aclweb.org/anthology/W19-7717/ (date of access: 01.07.2021).
4. *Nivre J., de Marneffe M.-C., Ginter F., Hajič J., Manning C., Pyysalo S., Schuster S., Tyers F., Zeman D.* (2020), Universal Dependencies v2: An Evergrowing Multilingual

Treebank Collection. In: Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), ELRA, Marseille, France, pp. 4034–4043. URL: https://www.aclweb.org/anthology/2020.lrec-1.497/ (date of access: 01.07.2021).
5. *Nivre J., Marongiu P., Ginter F., Kanerva J., Montemagni S., Schuster S., Simi M.* (2018), Enhancing Universal Dependency Treebanks: A Case Study. In: Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), Bruxelles, Belgium, pp. 102–107. URL: https://www.aclweb.org/anthology/W18-6012 (date of access: 01.07.2021).
6. *Schuster S., Manning C.* (2016), Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), ELRA, Portorož, Slovenia, pp. 2371–2378. URL: https://www.aclweb.org/anthology/L16-1376/ (date of access: 01.07.2021).
7. *Zeman D., Hajič J., Popel M., Potthast M., Straka M., Ginter F., Nivre J., Petrov S.* (2018), CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Bruxelles, Belgium, pp. 1–21. URL: https://www.aclweb.org/anthology/K18-2001/ (date of access: 01.07.2021).

Daniel Zeman
Charles University, Faculty of Mathematics and Physics, ÚFAL (Czech Republic)
E-mail: zeman@ufal.mff.cuni.cz

С. О. Савчук

S. O. Savchuk

НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА В ЗЕРКАЛЕ СТАТИСТИКИ[1]

THE RUSSIAN NATIONAL CORPUS THROUGH THE PRISM OF STATISTICS

Аннотация. Основными критериями оценки текстовых корпусов, принятыми в корпусной лингвистике, являются объем, репрезентативность, сбалансированность. Особенностью НКРЯ как большого представительного корпуса является то, что за все время его существования рост объема осуществлялся пропорционально, с соблюдением баланса в функциональном разнообразии текстов. В связи с планами дальнейшего развития корпуса был проведен масштабный анализ состава текстов НКРЯ по основным параметрам, которые традиционно используются для оценки репрезентативности корпусов. Среди них дата создания текстов, сфера функционирования, жанр и тематика текстов. В статье представлены результаты анализа статистической информации об основных метатекстовых параметрах корпуса.
Ключевые слова. Национальный корпус русского языка, состав текстов, репрезентативность, сбалансированность.

Abstract. The main criteria for evaluating text corpora accepted in corpus linguistics are corpus size, representativeness, and balance. A peculiarity of the RNC design as a large representative corpus is that its size increasing was proportional, while maintaining a balance in the functional diversity of texts. A large-scale analysis of text composition of the RNC was carried out using the parameters traditionally applied for evaluating corpora representativeness. Among them are date of text creation, functional sphere, genre, and text topic. The article presents the results of the analysis of statistical information about the main metatextual parameters of the RNC.
Keywords. The Russian National Corpus, corpus composition, representativeness, balance.

1. Введение

Национальный корпус русского языка проектировался в начале 2000-х годов как корпус современного русского языка по образцу больших национальных корпусов объемом 100 млн словоупотреблений. Это особый тип большого представительного корпуса, отражающий функционирование национального языка в современный период. За 20 лет существования ядро корпуса увеличилось в объеме и обросло системой специальных корпусов, но при этом, согласно статистике запросов, основной корпус письменных текстов остается самым вос-

[1] Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации в рамках Соглашения № 075-15-2020-793.