https://doi.org/10.5281/zenodo.5746834
УДК 81.33

Кузнецов А.В.

Кузнецов Алексей Валерьевич, кандидат исторических наук, научный сотрудник, Институт все общей истории РАН, Россия, 119334, г. Москва, Ленинский проспект, 32 а. E-mail: historyras@gmail.com.

Проект Universal Dependences и его вклад в развитии корпусных технологий: на примере аннотированных корпусов для латинского языка

Аннотация. За последние годы существенно выросло количество крупных языковых корпусов, что привело к растущему интересу к эмпирическим исследованиям в области лингвистики и корпусным методам. Одна из разновидностей лингвистических корпусов – трибанки – большие коллекции морфологически и синтаксически проанализированных предложений стали ценным ресурсом не только для традиционных филологических исследований, но и для задач компьютерной лингвистики, таких как автоматический морфологический и синтаксический анализ. Статья посвящена вкладу международного проекта Универсальные зависимости в развитие корпусных методов и инструментов анализа естественного языка на примере латиноязычных трибанков. Проведено сравнение состава латиноязычных корпусов. Рассмотрены наиболее универсальных инструменты обработки естественных языков и анализа текстов, использующих эти корпуса.
Ключевые слова: Универсальные зависимости, лингвистический корпус, трибанк, корпусное исследование, компьютерная лингвистика, обработка естественного языка, интеллектуальный анализ текстов, латинский язык.

Kuznetsov A.V.

Kuznetsov Alexey Valerevich, researcher, Institute of World History of Russian Academy of Sciences, Russia, 119334, Moscow, Leninskij prospekt, 32 a. E-mail: historyras@gmail.com.

The Universal Dependences Project and its Contribution to the Development of Corpus Technologies: An Example of Annotated Corpora for Latin

Abstract. The last years have seen a growing number of large natural language corpora leading to an increasing interest in empirical issues within the field of linguistics and the corpus-based methods. One type of linguistic corpora – treebanks – large collections of syntactically parsed sentences have recently emerged as a valuable resource not only for traditional philological researches, but for computational tasks such as automatic morphological and syntactical parsing. The article is devoted to the contribution of the international project Universal Dependencies to the development of corpus methods and tools for natural language processing by the example of Latin treebanks. The article compares the composition of the Latin corpora. The most universal tools for natural language processing and text analysis that these treebanks use are considered.
***Key words*:** Universal Dependencies, linguistic corpus, treebank, corpus-based research, computational linguistics, natural language processing, text mining, Latin.

Одной из тенденций в филологических исследованиях в настоящее время является всё большее использование корпусных методов. Согласно наиболее общему определению под лингвистическим корпусом понимается «представленный в электронном виде унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных языковых задач» [1, С. 3]. Одной из разновидностей лингвистических корпусов являются так называемые древовидные банки или трибанки. Древовидный банк – это текстовый корпус с многослойной аннотацией, в котором каждое слово помимо указания на его лемму (словарную форму), снабжено метками (тегами) с информацией о морфологии слова и синтаксических отношениях слов в предложении. Наличие такой многослойной разметки значительно увеличивает полезность корпусов. Лемматизация позволяет выполнять запросы для всех словоформ, подпадающих под одну лемму. Части речи и морфологические теги дают возможность выполнять запросы для конкретных комбинаций лингвистических функций на уровне слова, без необходимости ссылаться на словоформу. Синтаксическая маркировка позволяет искать группы слов, которые синтаксически связаны, независимо от их положения в тексте. Название – древовидный банк – объясняется тем, что синтаксическая структура предложения обычно представляется в виде древовидного графа, в котором словоформы являются узлами, а ребра выражают связь между словоформами.

В настоящей статье мы на примере трибанков латинского языка покажем историю развития проекта Универсальные зависимости, его роль для филологии, компьютерной лингвистики, а также других областей гуманитарных исследований.

Универсальные зависимости (англ. Universal Dependencies, https://universaldependencies.org/) – это международный открытый проект, направленный на разработку универсальной, кросслингвистической схемы разметки трибанков для большого количества языков, с целью унификация межъязыковой лингвистической типологии, упрощение кросслингвистических исследований, создание основы для разработки автоматизированных систем обработки естественного языка в том числе для многоязычных текстов [16].

Первой попыткой создания универсальной схемы аннотаций был проект Universal Dependency Treebank [17] в рамках которого были созданы корпуса для 6 языков в 2013 году и 11 языков в 2014 году. Вторым проектом стал HamleDT [5] по аннотированию корпусов уже для 30 языков в 2014 году.

Проект Универсальные зависимости ведет свое начало с 2014 года. Разработанная в его рамках схема разметки языковых корпусов стала плодом объединения предыдущих наработок в сфере создания аннотированных корпусов: универсальных тегов частеречной разметки Google (Google Universal Part-of-Speech Tags), универсальных стэнфордских зависимостей (Universal Stanford Dependencies) и средства преобразования различных наборов тегов Interset interlingua [16]. Общий теоретический подход проекта состоит в том, чтобы разработать универсальный перечень меток разметки и руководящих принципов для облегчения согласованной аннотации схожих конструкций на разных языках, при необходимости допуская расширения для конкретных языков.

Разметка корпусов проекта Универсальные зависимости содержит три набора признаков: теги частей речи (универсальный набор из 17 тегов частей речи), морфологические признаки (лемма, теги лексических признаков и теги флективных признаков), синтаксические зависимости (более 40 вариантов синтаксических зависимостей). Подробно последняя схема ан-

нотации описана в статье Й. Нивре и др. [16].

Схема разметки корпусов проекта Универсальные зависимости стала фактически стандартом для кросс - лингвистически сопоставимых морфологических и синтаксических разметок. Проект весьма успешен и стремительно развивается: всего за семь лет он вырос с 10 аннотированных корпусов для 10 языков и дюжины исследователей в 2014 году до 202 корпусов для 114 языков (на момент написания статьи), созданных при участии более 400 исследователей со всего мира в нынешнем году [15]. Примерно каждые полгода анонсируются новые релизы проекта. Особенностью проекта является то, что помимо корпусов современных языков, множество исследователей создают трибанки для очень редких и малоресурсных языков (африканские языки, языки коренных народов Америки и т.п.), а также исторических и древних языков, таких как латинский, древнегреческий, древнерусский, аккадский, готский, коптский, старофранцузский, древнеирландский, старотурецкий и др.

Корпуса для латинского языка появились в одном из первых релизов – версии 1.2 в конце 2015 года. В настоящее время для латинского языка в проекте Универсальные зависимости доступны уже пять трибанков (Табл. 1):

Perseus Treebank сконвертирован из Ancient Greek and Latin Dependency Treebank (AGLDT) (https://perseusdl.github.io/treebank_data/) – самого раннего трибанка для латинского и древнегреческого языков. Проект был начат в 2006 году в Университете Тафтса. В настоящее время он разрабатывается и поддерживается совместными усилиями сотрудников Университета Тафтса и Университета Лейпцига [4]. Трибанк составлен на основе классических и позднеантичных латинских текстов, включает 29 138 слов и 2 273 предложения.

PROIEL Project Treebank создан в Университете Осло в рамках проекта PROIEL (Pragmatic Resources in Old Indo-European Languages) – корпуса классических и позднеантичных латинских и древнегреческого текстов, а также параллельных текстов Нового завета на древнегреческом языке и его самых ранних переводов на индоевропейские языки: латынь, готский, старославянский и классический армянский (http://syntacticus.org/). Трибанк сформирован на основе частичной выборки из классических и позднеантичных латинских текстов, включает 200 163 слова и 18 411 предложений [7].

Index Thomisticus Treebank (IT-TB) (https://itreebank.marginalia.it/) основан на корпусе текстов Index Thomisticus (https://www.corpusthomisticum.org/) – старейшего проекта в области компьютерной лингвистики и цифровой гуманитаристики, он объединяет 118 сочинений Фомы Аквинского и 61 текст авторов его круга, всего более 11 миллионов слов. Index Thomisticus Treebank самый объемный из существующих трибанков для латинского языка. Он включает 1, 2 и 3 книги из Summa contra Gentiles, отрывки из Summa Theologiae Фомы Аквинского и Scriptum super Sententiis Magistri Petri Lombardi [10]. В 2020 году трибанк содержал 353 035 слов и 21 011 предложений, в последней версии 2.8 он содержит уже 450 515 слов и 26 977 предложений. Подобный объем сопоставим с объемом аннотированных корпусов для современных языков.

В 2020 году в релизе 2.6 в проекте появился четвертый латиноязычний трибанк – Late Latin Charter Treebank (LLCT), составленный на основе 521 раннесредневекового частноправового юридического документа (хартии), написанных в Тоскане между 774 и 897 годами для регистрации частных сделок, таких как продажа, обмен и сдача внаем собственности [5]. Язык этих документов представляет собой нестандартную разновидность латыни, отличную как от классической, так и от средневековой латыни с точки зрения правописания, морфологии и синтаксиса.

Трибанк содержит 257 918 слов и 9 023 предложения.

В 2021 году был добавлен уже пятый на данный момент корпус UDante, основанный на латинских текстах Данте Алигьери, взятых из корпуса DanteSearch (https://dantesearch.dantenetwork.it), созданного в Пизанском университете [14]. Трибанк содержит 55 697 слов и 1 721 предложение. Готовые лингвистические модели для этого трибанка на момент написания тезисов ещё не доступны.

Таблица 1. Латиноязычные трибанки проекта Universal Dependencies.

Автор	Произведение	Время создания	Жанр
Perseus Treebank			
Август	Res Gestae Divi Augusti	I век н.э.	Автобиография, историография.
Цезарь	Commentarii de Bello Gallico	I век до н.э.	Историческое произведение
Цицерон	In Catilinam	I век до н.э.	Риторическое произведения
Иероним	Vulgata	V век н.э.	Религиозное произведение
Вергилий	Aeneid0	I век до н.э.	Эпос
Овидий	Metamorphoses	I век до н.э.	Эпос
Петроний	Satyricon	I век н.э.	Новелла
Федр	Fabulae	I век н.э.	Басня
Пропертий	Elegiae	I век до н.э.	Элегия
Саллюстий	Bellum Catilinae	I век до н.э.	Историческое произведение
Светоний	De vita Caesarum	II век н.э.	Историческое произведение
Тацит	Historiae	II век н.э.	Историческое произведение
Всего слов: 29 138			
PROIEL Project Treebank			
Цезарь	Commentarii de Bello Gallico	I век до н.э.	Историческое произведение
Цицерона	Epistulae ad Atticum, De officiis	I век до н.э.	Риторические произведения
Иероним	Vulgata	V век н.э	Религиозное произведение
Всего слов: 200 163			
Index Thomisticus Treebank			
Фома Аквинский	Summa contra Gentiles, Scriptum super Sententiis Magistri Petri Lombardi, Summa Theologiae	XIII век	Теологические трактаты
Всего слов: 353 035			
Late Latin Charter Treebank			
---	Раннесредневековые хартии	774 - 897 гг.	Юридические документы.
Всего слов: 257 918			
UDante			
Данте Алигьери	De vulgari eloquentia, Monarchia,Letters, Questio de aqua et terra, Eclogues	XIV век	Литературные произведения разных жанров: риторические, политические, технические, философские, корреспонденция.
Всего слов: 55 697			

Аннотированные корпуса предоставляют обширный эмпирический материал для анализа лексики и грамматики, особенностей употребления тех или иных морфологических форм слов, употребления тем или иным автором определенных грамматических конструкций, диахронических исследований языка и многого другого. Помимо этого корпуса выступают в качестве обучающего набора данных в задачах машинного обучения и создании программных продуктов для обработки естественного языка.

С точки зрения практики обработка естественного языка может быть разбита на несколько этапов. Начальным этапом в любом случае будет предварительная обработка текста, с целью преобразования необработанного текста в набор данных, пригодный для анализа. Предварительная обработка может включать в различном сочетании следующие операции [2, P. 45-59]: 1. Токенизация – разбиение текста на фрагменты. 2. Очистка текста – удаление пустых строк, чисел, знаков препинания, типографских знаков. 3. Удаление стоп-слов – малозначимых и низкоинформативных (служебные части речи, местоимения, числительные и др.). 4. Лемматизация – приведение слова к словарной форме. 5. Частеречная разметка – морфологический анализ слов. 6. Синтаксический анализ предложений.

Лингвистические модели, обученные на основе почти всех древовидных банков проекта Универсальные зависимости сотрудниками Института формальной и прикладной лингвистики физико-математического факультета Карлова университета, дают возможность проводить предварительную обработку, в том числе морфологический разбор слов и синтаксический разбор предложений в неразмеченных текстах, что является базисом последующего анализа текста. Доступные в настоящее время лингвистические модели для латинского языка (версия 2.5, ноябрь 2019 года) демонстрируют хороший результат для лемматизации и частеречной разметки (Табл. 2).

Таблица 2. Сравнение качества работы моделей на основе латиноязычных трибанков (https://ufal.mff.cuni.cz/udpipe/models/).

Модель	Токенизация	Универсальная частеречная разметка	Специфическая частеречная разметка	Лемматизация
IT-TB	100.0%	97.1%	93.0%	98.0%
PROIEL	99.9%	94.5%	94.7%	94.5%
Latin-Perseus	100.0%	83.3%	67.2%	78.0%

Благодаря унификации разметки корпусов языковые модели, созданные в проекте Универсальные зависимости, используются в нескольких наиболее популярных программных продуктах, применимых к анализу в том числе и латиноязычных текстов:

Во-первых, UDPipe – бесплатных библиотек и пакеты на нескольких языках программирования: R, C++, Python, Perl, Java, C#, созданные в Институте формальной и прикладной лингвистики физико-математического факультета Карлова университета в Праге [12].

Во-вторых, Classical Language Toolkit (CLTK) [13] – специализированная библиотека на языке Python для обработки классических и древних языков. Разработка CLTK ведется с 2014 году. В настоящее время CLTK поддерживается множеством энтузиастов.

В-третьих, Stanza – новейшая разработка Стэнфордского университета, библиотека на языке Python, поддерживающая 66 языков, включая латынь [11].

Наконец, Orange – бесплатный инструмент для интеллектуального анализа данных, визуализации и построения моделей машинного обучения [8]. Orange создан в лаборатории биоинформатики на факультете компьютерных и информационных наук Университета Любляны, имеет графический интерфейс и не требует написания кода.

В рамках проекта Универсальные зависимости и с использованием лингвистических моделей, созданных на основе корпусов проекта, проводятся регулярные соревнования среди команд разработчиков программных продуктов. На таких соревнованиях опробуются различные подходы к решению задач обработки естественного языка. В 2020 году впервые прошло соревнование EvaLatin 2020, посвященное оценке инструментов обработки естественного языка для латыни [9]. Одной из целей проведения соревнований, по мысли организаторов, было стимулирование исследований в области компьютерных технологий для классических языков. В соревновании были предложены две задачи: лемматизация и частеречная разметка. Каждая задача имела по три подзадачи: 1. Классический вариант, когда тестовые данные относятся к тому же жанру и временному периоду, что и обучающие данные; 2. Кросс-жанровый вариант – тестовые данные относятся к другому жанру, но к тому же периоду времени; 3. Кросс-временной вариант – тестовые данные относятся к другому периоду времени, нежели обучающие данные. В качестве материала использовались тексты произведений Цезаря, Цицерона, Сенеки, Плиния Младшего и Тацита. Для кросс-жанрового теста были использованы поэтические тексты Горация. Для кросс-временного – тексты Фомы Аквинского. Выбранные тексты были аннотированы с помощью моделей UDPipe, а затем вручную выверены специалистами по латинскому языку. По причине ограничений во время пандемии соревнование не смогло пройти в очном формате, но команды разработчиков предоставили свои результаты. Лидерами в этот раз оказались разработчики UDPipe. Так в классическом тесте задачи лемматизации они добились точности 96,19 %, в кросс-жанровом – 87,13%, в кросс-временном 91,01%. В классическом тесте задачи частеречной разметки были достигнуты результаты точности 96,74 %, в кросс-жанровом – 91,11%, в кросс-временном 87,69% [9, P. 108-109]. Это очень хорошие показатели, дающие надежду на скорое появление ещё более совершенных инструментов анализа текстов на латинском языке.

Наконец нельзя не упомянуть, что размеченные корпуса используются в преподавании латинского языка. Так при морфологической и синтаксической разметке Latin Dependency Treebank применяется, в том числе, и такой подход, при котором разметку проводят студенты. Затем результат их трудов проверяет преподаватель. Данный подход опробовался в нескольких университетах США и Европы. Он показал, с одной стороны, хорошие результаты в изучении студентами сложных грамматических конструкций. С другой стороны, он дал возможность наглядно контролировать успеваемость, автоматически определяя сильные и слабые стороны отдельных учащихся [3, P. 546-549].

За последние десятилетия широкое распространение цифровых технологий увеличило доступность текстовых источников, радикально изменив повседневную жизнь ученых-гуманитариев, которые теперь могут обрабатывать обширные эмпирические данные ранее недоступными методами. Стремительное развитие получили и инструменты анализа текста, применимые не только к современным, но и к древним языкам. Существенную роль в этом процессе сыграл проект Универсальные зависимости, в рамках которого разработана универсальная схема аннотации корпусов для множества языков. Древо-

видные банки латинского языка за время реализации проекта Универсальные зависимости как увеличивались количественно, так и росли в объеме. В настоящее время в проекте представлены 5 корпусов включающие тексты, созданные с I века до нашей эры до XIV века нашей эры, написанные в разных жанрах.

СПИСОК ЛИТЕРАТУРЫ

1. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: уч. пос. 3-е изд., переработанное. СПб.: Издательство Санкт-Петербургского университета, 2020. 234 с.
2. Anandarajan M., Hill C., Nolan T. Practical Text Analytics. Maximizing the Value of Text Data. (Advances in Analytics and Data Science. Vol. 2.) Springer, 2019. 285 p.
3. Bamman D., Crane G. Corpus linguistics, treebanks and the reinvention of philology // INFORMATIK 2010. Service Science–Neue Perspektiven für die Informatik. Band 2. Bonn, 2010. P. 542-551.
4. Bamman D., Crane G. The Latin Dependency Treebank in a cultural heritage digital library // Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007). Prague, Czech Republic, 2007. P. 33-40.
5. Cecchini F.M., Korkiakangas T., Passarotti M. A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages // Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). Marseille, 2020. P. 933-942.
6. HamleDT 2.0: Thirty Dependency Treebanks Stanfordized / Rosa R. // Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation. Reykjavik,2014.P.2334-2341.
7. Haug D.T., Jøndal M.L. Creating a Parallel Treebank of the Old Indo-European Bible Translations // Proceedings of Language Technologies for Cultural Heritage Workshop. (LREC 2008.) Marrakech, 2008. P. 27-34.
8. Orange: Data Mining Toolbox in Python / Demsar J., Curk T., Erjavec A. et al. // Journal of Machine Learning Research. 2013. N 14(Aug) . P. 2349−2353.
9. Overview of the EvaLatin 2020 Evaluation Campaign / Sprugnoli R., Passarotti M., Cecchini F.M., Pellegrini M. // Proceedings of LT4HALA 2020 – 1st Workshop on Language Technologies for Historical and Ancient Languages. Marseille, 2020. P. 105-110.
10. Passarotti M. The Project of the Index Thomisticus Treebank // Digital Classical Philology. Berlin, Boston: De Gruyter Saur, 2019. P. 299-320.
11. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages / Qi P., Zhang Y., Zhang Y., Bolton J., Manning C.D. // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020. P. 101-108.
12. Straka M. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task // Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning. Brussels, 2018. P. 197-207.
13. The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages / Johnson K.P., Burns P.J., Stewart J., Cook T., Besnier C., Mattingly W.J.B. // Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. 2021. P. 20-29.
14. UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works / Cecchini F.M., Sprugnoli R., Moretti G., Passarotti M. // Proceedings of the Seventh Italian Conference on Computational Linguistics. CEUR Workshop Proceedings, 2020.Bologna, Italy,March 1-3,2021.P.1-7.
15. Universal Dependencies / de Marneffe M.-C., Manning C., Nivre J., Zeman D. // Computational Linguistics, 2021. V. 47. N 2. P. 255-308.
16. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection / Nivre J., Marneffe M.-C., Ginter F et al. // Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). Marseille, 11–16 May 2020. P. 4034-4043.
17. Universal Dependency Annotation for Multilingual Parsing / Ryan McDonald at al. // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, 2013. P. 92-97.

REFERENCES (TRANSLITERATED)

1. Zaharov V.P., Bogdanova S.Ju. Korpusnaja lingvistika: uch. pos. 3-e izd., pererabotannoe. SPb.: Izdatel'stvo Sankt-Peterburgskogo universiteta, 2020. 234 s.
2. Anandarajan M., Hill C., Nolan T. Practical Text Analytics. Maximizing the Value of Text Data. (Advances in Analytics and Data Science. Vol. 2.) Springer, 2019. 285 p.
3. Bamman D., Crane G. Corpus linguistics, treebanks and the reinvention of philology // INFORMATIK2010.Service Science–Neue Perspektiven für die Informatik.Band2.Bonn,2010.P.542-551
4. Bamman D., Crane G. The Latin Dependency Treebank in a cultural heritage digital library // Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007). Prague, Czech Republic, 2007. P. 33-40.
5. Cecchini F.M., Korkiakangas T., Passarotti M. A New Latin Treebank for Universal Dependen-cies: Charters between Ancient Latin and Romance Languages // Proceedings of the 12th Con-ference on Language Resources and Evaluation (LREC 2020). Marseille, 2020. P. 933-942.
6. HamleDT 2.0: Thirty Dependency Treebanks Stanfordized / Rosa R. // Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation.Reykja-vik,2014.P.2334-2341.
7. Haug D.T., Jøndal M.L. Creating a Parallel Treebank of the Old Indo-European Bible Transla-tions // Proceedings of Language Technologies for Cultural Heritage Workshop. (LREC 2008.) Marrakech, 2008. P. 27-34.
8. Orange: Data Mining Toolbox in Python / Demsar J., Curk T., Erjavec A. et al. // Journal of Ma-chine Learning Research. 2013. N 14(Aug) . P. 2349−2353.
9. Overview of the EvaLatin 2020 Evaluation Campaign / Sprugnoli R., Passarotti M., Cecchini F.M., Pellegrini M. // Proceedings of LT4HALA 2020 – 1st Workshop on Language Technolo-gies for Historical and Ancient Languages. Marseille, 2020. P. 105-110.
10. Passarotti M. The Project of the Index Thomisticus Treebank // Digital Classical Philology. Ber-lin, Boston: De Gruyter Saur, 2019. P. 299-320.
11. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages / Qi P., Zhang Y., Zhang Y., Bolton J., Manning C.D. // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020. P. 101-108.
12. Straka M. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task // Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning. Brussels, 2018. P. 197-207.
13. The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages / Johnson K.P., Burns P.J., Stewart J., Cook T., Besnier C., Mattingly W.J.B. // Proceedings of the Joint Confer-ence of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. 2021. P. 20-29.
14. UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works / Cecchini F.M., Sprugnoli R., Moretti G., Passarotti M. // Proceedings of the Seventh Italian Con-ference on Computational Linguistics. CEUR Workshop Proceedings,2020.Bologna,Ita-ly,March 1-3,2021.P.1-7.
15. Universal Dependencies / de Marneffe M.-C., Manning C., Nivre J., Zeman D. // Computational Linguistics, 2021. V. 47. N 2. P. 255-308.
16. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection / Nivre J., Marn-effe M.-C., Ginter F et al. // Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). Marseille, 11–16 May 2020. P. 4034-4043.
17. Universal Dependency Annotation for Multilingual Parsing / Ryan McDonald at al. // Proceed-ings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, 2013. P. 92-97.

Поступила в редакцию 18.11.2021.
Принята к публикации 21.11.2021.

Для цитирования:
Кузнецов А.В. Проект Universal Dependences и его вклад в развитии корпусных технологий: на примере аннотированных корпусов для латинского языка // Гуманитарный научный вестник. 2021. №11. С. 128-135. URL: http://naukavestnik.ru/doc/2021/11/Kuznetsov.pdf