СЕССИЯ 3.
ИНТЕЛЛЕКТУАЛЬНЫЕ ТЕХНОЛОГИИ В ГУМАНИТАРНОЙ СФЕРЕ И ОБРАЗОВАНИИ

Латиноязычные трибанки в гуманитарных исследованиях и преподавании языка

Кузнецов А.В.
Институт всеобщей истории Российской академии наук (ИВИ РАН),
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0003-4755-250X, e -mail: historyras@gmail.com

Бурное развитие информационных технологий в последние десятилетия привело к их проникновению во все сферы жизни. В полной мере оно затронуло область гуманитарных исследований, породив новое направление цифровой гуманитаристики, и, конечно, образования. Значимым нововведением для лингвистики и преподавания иностранных языков стало использование электронных языковых корпусов. 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]. Языковые корпуса не только стали фактически стандартным инструментом лингвистических исследований, но и привели к появлению новых методов и приемов в обучении иностранным языкам [2]. Одной из разновидностей лингвистических корпусов являются так называемые «трибанки» — коллекции морфологически и синтаксически размеченных предложений. Название объясняется тем, что синтаксическая структура предложений представлена в них в виде древовидных графов.

В настоящих тезисах мы проведем сравнение существующих латиноязычных трибанков, наиболее популярных инструментов обработки естественных языков и анализа текстов, использующие эти трибанки, а также расскажем об опыте применения трибанков в обучении латинскому языку.

Долгое время латинский язык, с точки зрения развития доступных для него инструментов анализа естественного языка, относился к так называемым малоресурсным, поскольку инструментарий компьютерного анализа латиноязычных текстов был весьма ограниченным. Ситуация начала существенно меняться в 2014 г., когда был презентован проект Universal Dependencies (https://universaldependencies.org/) — открытый международный проект по разработке универсального формата морфологической

и синтаксической разметки трибанков для большого количества языков. Проект объединил в себе лучшие достижения в области аннотирования языковых корпусов: универсальные стэнфордские зависимости (Universal Stanford Dependencies), универсальные теги частеречной разметки Google (Google Universal Part-of-Speech Tags) и средство преобразования различных наборов тегов Interset interlingua [3, p. 4034].

Проект оказался весьма успешным. В начале 2015 г. версия 1.0 включала всего 10 трибанков для 10 языков. В последней версии 2.6 (май 2020 г.) проект Universal Dependencies содержит уже 163 трибанка для 92 языков. Релизы новых версий происходят с частотой два раза в год. Помимо трибанков для современных языков в проекте Universal Dependencies разработаны трибанки для древнегреческого, латинского, аккадского, древнерусского, церковнославянского, готского и коптского языков.

Латинский язык появился в проекте Universal Dependencies в конце 2015 г. в релизе 1.2. Сейчас для латинского языка в рамках проекта доступны четыре трибанка.

Perseus Treebank сконвертирован из Latin Dependency Treebank (LDT) (https://perseusdl.github.io/treebank_data/), созданного совместными усилиями преимущественно сотрудников и студентов Университета Тафтса и Университета Лейпцига [4]. Проект был начат в 2006 г. 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), направленного на синтаксическую аннотацию старейших версий Нового Завета на индоевропейских языках: латинском, греческом, готском, армянском и старославянском (http://syntacticus.org/). Как и предыдущий, трибанк сформирован на основе частичной выборки из классических и позднеантичных латинских текстов, включает 200 163 слова и 18 411 предложений [5].

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, включающего полное собрание текстов Фомы Аквинского и авторов его круга, всего более 11 миллионов слов. Составление Index Thomisticus в конце 1940-х гг. начал теолог Роберто Буза. Ныне Index Thomisticus Treebank включает размеченные морфологически и синтаксически книги 1, 2 и 3 из Summa contra Gentiles, а также выдержки из Scriptum super Sententiis Magistri Petri Lombardi и Summa Theologiae Фомы Аквинского.[6]. Трибанк содержит 353 035 слов и 21 011 предложений.

В 2020 г. был анонсирован четвертый латиноязычний трибанк — Late Latin Charter Treebank (LLCT) — составленный на основе 521 ранне-

средневекового частноправового юридического документа (хартии), написанных в Тоскане между 774 и 897 годами для регистрации частных сделок, таких как продажа, обмен и сдача внаем собственности [7]. Трибанк содержит 257 918 слов. Готовые лингвистические модели для этого трибанка на момент написания тезисов еще не доступны.

Проект Universal Dependencies стал определенной вехой в создании не только размеченных лингвистических корпусов, но и инструментов обработки естественных языков (Natural Language Processing, NLP) — области искусственного интеллекта и математической лингвистики, направленной на изучение методов анализа и синтеза естественного языка. Обработка естественного языка сегодня применяется во многих сферах. В число ее наиболее известных прикладных задач входит машинный перевод текстов, информационный поиск, автоматическая классификация и кластеризация текстов, автоматическая аннотация и реферирование текстов, разработка рекомендательных и вопросо-ответных систем.

Трибанки являются необходимым составляющим при создании программных продуктов для обработки и анализа текста. Они выступают в качестве обучающей выборки, на основе которой строятся лингвистические модели. Пригодность трибанка для лингвистического анализа, если рассматривать его именно как обучающую выборку, оценивается по размеру и содержанию [8, p. 11]. Чем трибанк объемнее и чем более хронологически, жанрово и тематически его содержание совпадает с анализируемым текстом, тем выше будет качество анализа. Подробный перечень авторов и произведений, аннотированных в трибанках, дан в табл. 1.

Лингвистические модели, обученные на основе трибанков, дают возможность проводить морфологический разбор слов и синтаксический разбор предложений в неразмеченных текстах, что является базисом анализа текста. Качество работы лингвистических моделей Universal Dependencies весьма высокое (табл. 2).

Среди множества программных продуктов, применимых к анализу латиноязычных текстов, неполный перечень которых можно увидеть в [9], укажем в нашем обзоре три наиболее универсальные и, пожалуй, чаще всего используемые.

Во-первых, UDPipe — программное обеспечение, созданное в Институте формальной и прикладной лингвистики физико-математического факультета Карлова университета в Праге (https://ufal.mff.cuni.cz/udpipe/). UDPipe обладает широкими возможностями для обработки естественных языков и является наиболее универсальным, поскольку реализовано в виде бесплатных библиотек и пакетов на нескольких языках программирования: R, C++, Python, Perl, Java, C#.

Таблица 1

Состав латиноязычных трибанков 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	Aeneid	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			

Во-вторых, Classical Language Toolkit (CLTK) (https://cltk.org) — библиотека на языке Python для обработки классических и древних язы-

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%

ков. Разработка CLTK была начата в 2014 г. и в настоящее время поддерживается множеством энтузиастов.

В-третьих, Stanza — новейшая, анонсированная в 2020 г. библиотека на языке Python, разработанная в Стэнфордском университете (https://stanfordnlp.github.io/stanza/) и поддерживающая 66 языков, включая латынь. Пакет Stanza является надстройкой библиотеки PyTorch и работает с использованием компонентов нейронной сети.

Перечисленные программные продукты позволяют осуществлять с латиноязычными текстами все основные операции обработки естественного языка: токенизацию, лемматизацию, морфологический анализ, синтаксический анализ и др.

В сфере изучения латинского и других древних мертвых языков корпусные технологии применяются не так широко, как при изучении современных, но такой опыт имеется, он связан с особенностью разработки проекта Latin Dependency Treebank. При разметке используется три метода. В первом случае, разметка делается единолично хорошо подготовленным специалистом. Во втором, предложения независимо размечаются двумя специалистами, после чего их результат согласует третий. Наконец, аннотация проводится студентами, после чего результат проверяется преподавателем [10, p. 546—549]. Необходимо отметить, что разметка корпуса — это фактически морфологический и синтаксический разбор предложений, что является, вероятно, ведущим навыком специалистов по классической филологии. Аннотирование трибанков практикуется на курсах изучения классической филологии в шести университетах США и Италии (Университет Тафтса, Университет Брандейс, Университет Фурмана, Университет Миссури в Канзас-Сити и Университет Небраски в Линкольне, Колледж Святого Креста) Данный метод, с одной стороны, показал 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 [10, p. 546—548.].

Латинские трибанки — это не просто лингвистические базы данных, но основа для филологических исследований и создания инструментов анализа текста. Уже сейчас существующие трибанки в совокупности содержат внушительное число аннотированных предложений из произведений различных жанров и эпох. За несколько прошедших лет они продемонстрировали стремительный рост. Также стремительно развиваются инструменты автоматического анализа текстов. Во многом это связанно с развитием проекта Universal Dependencies.

Литература

1. *Захаров В.П., Богданова С.Ю.* Корпусная лингвистика: учеб. пособие. 2-е изд. СПб.: СПбГУ. РИО. Филологический факультет, 2013.
2. *Горина О.Г.* Инструменты корпусного анализа в обучении иностранному языку // Вестник Томского государственного университета. 2018. № 435. С. 187—194.
3. *Nivre J., de Marneffe M.-C., Ginter F, Hajivc J, Manning C. D., Pyysalo S., Schuster S., Tyers F., Zeman D.* Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection // Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). Marseille, 11—16 May 2020. P. 4034—4043.
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. *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.
6. *Passarotti M.* The Project of the Index Thomisticus Treebank // Digital Classical Philology. Berlin, Boston: De Gruyter Saur, 2019. P. 299—320.
7. *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.
8. *McGillivray B.* Methods in Latin Computational Linguistics. Brill: Leiden, Boston, 2014.
9. *Burns P.J.* Building a Text Analysis Pipeline for Classical Languages // Digital Classical Philology, Ancient Greek and Latin in the Digital Revolution / Ed. Berti M. De Gruyter (Berlin), 2019. P. 159—176.
10. *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.

Сведения об авторе

Кузнецов Алексей Валерьевич, кандидат исторических наук, научный сотрудник, Институт всеобщей истории Российской академии наук (ИВИ РАН), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0003-4755-250X, e-mail: historyras@gmail.com