УДК 81.33
DOI:10.17853/2587-6910-2021-04-61-65

ДРЕВНЕГРЕЧЕСКИЕ АННОТИРОВАННЫЕ КОРПУСА В ЦИФРОВОЙ ГУМАНИТАРИСТИКЕ И ПРЕПОДАВАНИИ ЯЗЫКА

ANCIENT GREEK ANNOTATED CORPORA FOR DIGITAL HUMANITIES AND LANGUAGE TEACHING

Алексей Валерьевич Кузнецов
научный сотрудник
historyras@gmail.com
Институт всеобщей истории РАН, Москва, Россия

Alexey Valer'evich Kuznetsov
Institute of World History of Russian Academy of Sciences, Moscow, Russia

Аннотация. *Корпусные методы становятся все более значимыми в современных исследованиях в исторической лингвистике и в цифровых гуманитарных науках. Одна из разновидностей лингвистических корпусов — трибанки (большие коллекции синтаксически проанализированных предложений) стали ценным ресурсом не только для традиционных лингвистических и филологических исследований, но и для задач компьютерной лингвистики, таких как автоматический морфологический и синтаксический анализ. Статья посвящена сравнению трибанков древнегреческого языка, рассмотрению наиболее универсальных инструментов обработки естественных языков и анализа текстов, использующих эти трибанки, а также опыту применения трибанков в обучении древнегреческому языку.*

Ключевые слова: *корпусная лингвистика, трибанки, древнегреческий язык, цифровая гуманитаристика, обработка естественного языка.*

Abstract. *The corpus-based methods are becoming increasingly central to present-day research in historical linguistics and digital humanities. One type of linguistic corpora — treebanks (large collections of syntactically parsed sentences) have recently emerged as a valuable resource not only for traditional linguistic and philological researches, but for computational tasks such as automatic morphological and syntactical parsing. The article is devoted to the comparison of ancient Greek treebanks, the most universal tools for natural language processing and text analysis that these treebanks use are considered, and a description of the experience of using treebanks in teaching the ancient Greek language is given.*

Keywords: *corpus linguistic, treebanks, ancient Greek, digital humanities, natural language processing.*

В настоящее время информационные технологии продолжают все глубже проникать во все сферы жизни. В области гуманитарных исследований это привело к появлению нового направления — цифровых гуманитарных наук (англ. *digital humanities*), объединяющих методику традиционных гуманитарных наук с компьютерными науками, расширяя базу и инструментарий исследований, открывая новые возможности для сбора, анализа и визуализа-

ции данных. Доступность большого количества цифровых текстов, относящихся к различным эпохам, предопределила рост роли лингвистических корпусов, корпусных методов и технологий автоматической обработки текстов на естественных языках, прежде всего для лингвистики, истории и преподавания иностранных языков [1]. 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» [2, с. 3]. Одной из разновидностей лингвистических корпусов являются так называемые трибанки или древовидные банки — это текстовые корпуса, в которых каждое слово снабжено аннотациями с информацией о морфологии и синтаксических отношениях слов в предложении. Название объясняется тем, что наиболее распространенным способом представления синтаксической структуры предложений являются древовидные графы.

Можно выделить три основных направления применения трибанков:

1. Они являются источником данных для проведения исследований лексики и грамматики, анализа употребления языковых элементов и конструкций.

2. Трибанки служат основой для создания компьютерных моделей языка, разработки, настройки и тестирования автоматизированных систем машинного перевода, распознавания речи, информационного поиска и других задач машинного обучения.

3. Они используются при изучении языков.

В статье мы проведем сравнение существующих синтаксически аннотированных корпусов древнегреческого языка, наиболее универсальных и популярных инструментов обработки естественных языков и анализа текстов, использующих эти трибанки, а также опишем практические аспекты применения трибанков в обучении древнегреческого языку.

Древнегреческий язык, характеризующийся богатой морфологией и свободным порядком слов, оказался довольно трудным для разработки корпусных инструментов и программ для обработки естественных языков. Стандартные приложения для анализа текстов, разработанные для современных языков, преимущественно английского, были мало применимы для древнегреческого. Но, не смотря на все трудности, в настоящее время известно девять аннотированных трибанков, созданных в рамках разных проектов [3; 4, p. 109–110]. В числе этих корпусов наиболее известны и представляют особый интерес два: Ancient Greek Dependency Treebank и PROIEL Project трибанк.

Ancient Greek Dependency Treebank от Perseus Project (https://perseusdl.github.io/treebank_data/) начали составлять в 2006 г. в Университете Тафтса, ныне поддерживается преимущественно сотрудниками Университета Лейпцига [5]. Трибанк основан на архаических, классических и постклассических поэтических и прозаических текстах, включает 202 989 слов и 13 919 предложений.

PROIEL Project Treebank создан в Университете Осло в рамках проекта PROIEL (Pragmatic Resources in Old Indo-European Languages), цель которого — синтаксическая и морфологическая аннотация старейших версий Нового Завета на индоевропейских языках: греческом, латинском, готском, армянском и старославянском (http://syntacticus.org/). Сформирован на основе выборки из текстов Нового Завета, «Истории» Геродота и «Хроники» Георгия Сфрандзи, включает 213 999 слов и 17 080 предложений [6].

Успех данных трибанков можно объяснить тем, что их разработчики вышли за рамки собственных исследований и подключились к проекту «Универсальные зависимости». «Универсальные зависимости» (Universal Dependencies, https://universaldependencies.org/) — это международный проект, направленный на разработку универсальной, кросс-лингвистической формы 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 [7, p. 4034]. Разработка универсального формата разметки помогает сравнивать данные на разных языках, а также разрабатывать инструменты анализа текстов,

успешно применимые к текстам на разных языках, включая и многоязычные тексты.

Проект «Универсальные зависимости» был инициирован в 2014 г., оказался весьма успешным и сейчас стремительно развивается. Если первая версия в начале 2015 г. объединяла 10 трибанков для 10 языков, то в последней версии 2.7 (ноябрь 2020 г.) проект «Универсальные зависимости» содержит уже 183 трибанка для 104 языков. Релизы новых версий появляются примерно каждые полгода. Помимо трибанков для современных языков в рамках проекта «Универсальные зависимости» разрабатываются аннотированные корпуса для таких древних языков, как древнегреческий, латинский, аккадский, древнерусский, готский, коптский и др. Древнегреческий язык появился в проекте «Универсальные зависимости» в конце 2015 г. в релизе 1.2.

Аннотированные корпуса содержат информацию, которая может быть использована для количественного анализа различных явлений в древних текстах с высокой степенью точности. Например, с помощью трибанков можно узнать, какие глаголы чаще встречаются с существительными мужского рода, а какие с существительными женского рода, или насколько один автор использует более сложные грамматические конструкции, чем другой. Помимо этого трибанки выступают в качестве обучающей выборки под различные задачи машинного обучения. Они являются необходимым компонентом при обучении языковых моделей и создании программных продуктов для обработки и анализа текста.

Если посмотреть на анализ текста с практической точки зрения, то его можно разделить на несколько этапов. В подавляющем большинстве случаев одним из начальных этапов будет предварительная обработка текста, цель которой состоит в преобразовании текста в набор данных, пригодный для анализа. Предварительная обработка может включать в различном сочетании следующие операции [8, p. 45–59]: 1. Токенизация — разбиение текста на фрагменты. 2. Очистка текста — удаление лишних пробелов и пустых строк, типографских знаков, чисел, знаков препинания, перевод всех букв в нижний регистр. 3. Удаление стоп-слов — малозначимых и низкоинформативных (как правило, служебные части речи, местоимения, числительные). 4. Лемматизация — приведение слова к словарной форме. 5. Частеречная разметка — морфологический анализ слов. 6. Синтаксический парсинг — синтаксический анализ предложений. Предварительная обработка является основой для дальнейших шагов в анализе текста.

Лингвистические модели, созданные на основе трибанков, дают возможность проводить предварительную обработку, морфологический разбор слов и синтаксический разбор предложений в неразмеченных текстах, что является базисом анализа текста. Доступные в настоящее время в рамках проекта «Универсальные зависимости» модели для древнегреческого языка (версия 2.5, ноябрь 2019 г.) демонстрируют вполне приемлемый результат для лемматизации и частеречной разметки (табл. 1). Но качество моделей и инструментов анализа постоянно растет. Недавно удалось получить очень высокие результаты в морфологическом анализе (точность около 95 %) и лемматизации (точность около 99 %) на основе корпуса текстов с акцентом на греческих папирусах [9].

Среди программных продуктов, применимых к анализу древнегреческих текстов, укажем в нашем обзоре три наиболее универсальные и 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/). Оно дает возможность пользоваться готовыми моделями или тренировать собственные для широкого круга задач обработки естественного языка. UDPipe доступно в виде бесплатных библиотек

Таблица 1

Сравнение качества работы моделей на основе древнегреческих трибанков, % (https://ufal.mff.cuni.cz/udpipe/models/)

Модель	Токенизация	Универсальная частеречная разметка	Специфическая частеречная разметка	Лемматизация
Ancient Greek-Perseus	100,0	82,2	72,2	82,7
Ancient Greek-PROIEL	100,0	96,0	96,2	93,2

и пакетов на разных языках программирования: R, C++, Python, Perl, Java, C#.

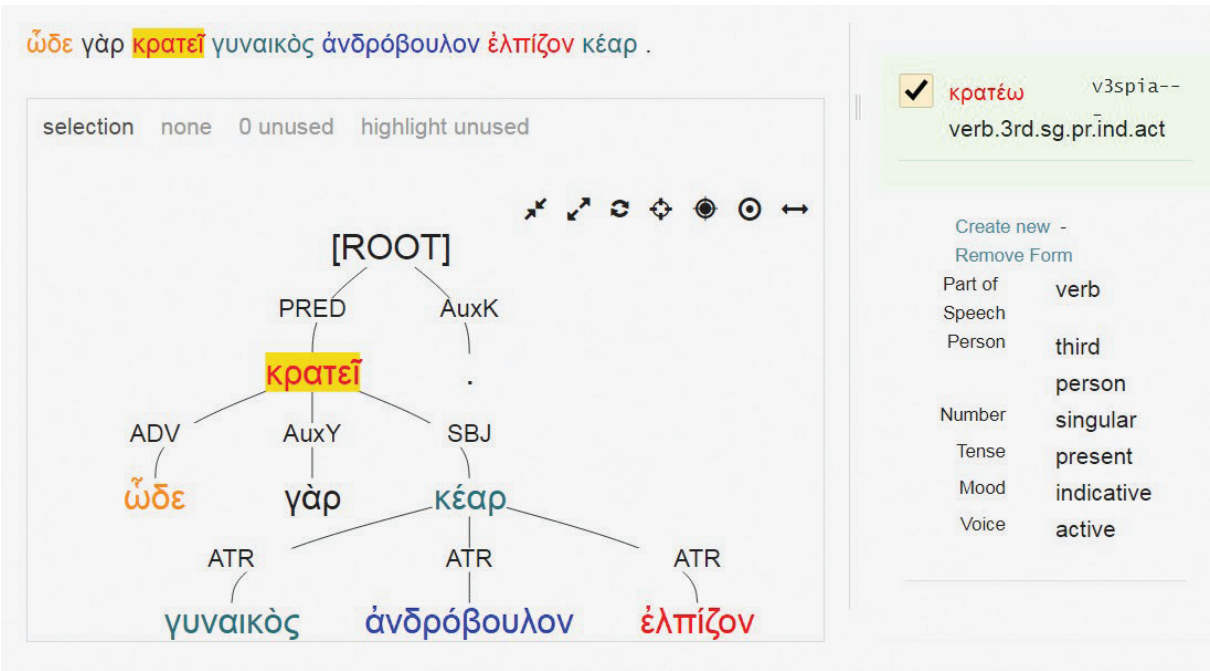
Во-вторых, Classical Language Toolkit (CLTK) (https://cltk.org) — фреймворк на языке Python для обработки древних, классических и средневековых языков: египетского иероглифического, древнегреческого, латыни, арамейского, санскрита, классического китайского и др. Помимо всего CLTK включает в себя обширные коллекции текстов на поддерживаемых языках. Разработка фреймворка началась в 2014 г. и в настоящее время поддерживается множеством энтузиастов.

В-третьих, Stanza — новейшая разработка Стэнфордского университета, надстройка библиотеки PyTorch на языке Python, работающая с использованием компонентов нейронной сети (https://stanfordnlp.github.io/stanza/). Stanza включает в себя предварительно обученные модели для 66 языков, в их числе и древнегреческий.

Перечисленные программные продукты позволяют осуществлять с текстами на древнегреческом языке все основные операции обработки естественного языка: токенизацию, лемматизацию, частеречную разметку, синтаксический парсинг и др.

Переходя к вопросу применения трибанков в обучении языков, следует указать, что первые попытки использования языковых корпусов в преподавании относятся уже к началу 1980-х гг. С развитием вычислительной техники и появлением большого количества электронных корпусов их роль в преподавании языков все более возрастает. Можно выделить несколько основных областей применения языковых корпусов в образовании: 1. Данные, полученные в ходе изучения национальных корпусов, повлияли на подход к составлению словарей, справочников, учебных пособий и курсов иностранных языков. 2. Материалы языковых корпусов становятся источником разнообразных дидактических материалов, учебных заданий и основой самостоятельных проектов. 3. При обучении профессиональному переводу параллельные корпуса позволяют увидеть определенные закономерности и лингвистические законы в тексте оригинала и тексте перевода.

В сфере изучения древнегреческого и других древних мертвых языков корпусные технологии применяются не так широко, как при изучении современных, но и здесь трибанки показали свою полезность. Во-первых, уже аннотированные предложения можно использовать для визуализации сложных примеров синтаксиса. Во-вторых, участие студентов в разметке предложений для корпусов является прекрасным упражнением по синтаксическому и морфологическому анализу текстов. При составлении Ancient Greek Dependency Treebank проекта Perseus используются три метода разметки. В первом случае разметка делается самосто-



Редактор Alpheios для аннотации трибанков

ятельно хорошо подготовленным специалистом. Во втором – предложения независимо размечаются двумя специалистами, после чего их результат согласует третий. И в третьем случае аннотация выполняется студентами, после чего преподаватель проверяет результат [10, p. 546–549]. Аннотирование трибанков практикуется на курсах изучения классической филологии в университетах США и Италии. Данный метод, с одной стороны, показал хорошие результаты для понимания учащимися древнегреческой лексики (определение лемм и частей речи), грамматики (морфологический анализ словоформ) и синтаксиса (выявление зависимости между словами и фразами в предложении). С другой стороны, он позволяет наглядно контролировать успеваемость, определяя сильные и слабые стороны отдельных учащихся [10, p. 546–548]. Разметка трибанков выполняется в визуальном редакторе Alpheios на онлайн-платформе Perseids (https://sosol.perseids.org/sosol/user/user_dashboard) (рисунок). Морфологическая форма каждого слова описывается в правой части редактора, а синтаксические отношения визуально отображаются в виде графа со сказуемым в его вершине.

Корпусные технологии существенно расширяют исследовательские возможности гуманитариев всех специальностей. Трибанки древнегреческого языка — это не просто лингвистические базы данных, но и основа для филологических исследований и создания инструментов анализа текста. За несколько прошедших лет как трибанки, так и инструменты автоматического анализа текстов получили стремительное развитие. Во многом это связанно с развитием проекта «Универсальные зависимости» и стремлением к унификации разметки лингвистических корпусов.

Список литературы

1. *Горина, О. Г.* Инструменты корпусного анализа в обучении иностранному языку / О. Г. Горина. Текст: непосредственный // Вестник Томского государственного университета. 2018. № 435. С. 187–194.
2. *Захаров, В. П.* Корпусная лингвистика: учебное пособие / В. П. Захаров, С. Ю. Богданова. 2-е изд. Санкт-Петербург: Изд-во С.-Петерб. гос. ун-та, 2013. С. 3. Текст: непосредственный.
3. *Robie, J.* Nine Kinds of Ancient Greek Treebanks. Open Data for Digital Biblical Humanities / J. Robie. URL: http://jonathanrobie.biblicalhumanities.org/blog/2017/12/20/treebanks-for-ancient-greek. Text: electronic.
4. *Keersmaekers, A.* Creating, Enriching and Valorizing Treebanks of Ancient Greek / A. Keersmaekers, W. Mercelis, C. Swaelens, T. Van Hal. Text: print // Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019). Paris, 2019. P. 109–117.
5. *Bamman, D.* The Latin Dependency Treebank in a cultural heritage digital library / D. Bamman, G. Crane. Text: print // Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007). Prague: Czech Republic, 2007. P. 33–40.
6. *Haug, D. T.* Creating a Parallel Treebank of the Old Indo-European Bible Translations / D. T. Haug, M. L. Jondal. Text: print // Proceedings of Language Technologies for Cultural Heritage Workshop. (LREC 2008.) Marrakech, 2008. P. 27–34.
7. *Nivre, J.* Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection / J. Nivre, M.-C. de Marneffe, F. Ginter [and others]. Text: print // Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). Marseille, 11–16 May. 2020. P. 4034–4043.
8. *Anandarajan, M.* Practical Text Analytics. Maximizing the Value of Text Data / M. Anandarajan, C. Hill, T. Nolan. Advances in Analytics and Data Science. Springer, 2019. Vol. 2. Text: print.
9. *Keersmaekers, A.* Creating a richly annotated corpus of papyrological Greek: The possibilities of natural language processing approaches to a highly inflected historical language / A. Keersmaekers. Text: print // Digital Scholarship in the Humanities, Vol. 35, Issue 1. April 2020. P. 67–82.
10. *Bamman, D.* Corpus linguistics, treebanks and the reinvention of philology / D. Bamman, G. Crane. Text: print // INFORMATIK 2010. Service Science–Neue Perspektiven für die Informatik. Band 2. Bonn, 2010. P. 542–551.