

Обзор корпуса латинского языка LatinISE Corpus

Адрес ресурса: <https://www.sketchengine.eu/latinise-corpus/>

Проектная работа, письменная часть

Иванов Даниил
Малкина Екатерина
Кротова Алина
Чиркин Андрей
Жильцов Даниил

Проект подготовили:

dkivanov@edu.hse.ru
edmalkina@edu.hse.ru
aakrotova@edu.hse.ru
adchirkin@edu.hse.ru
dazhiltsov@edu.hse.ru

Составители ресурса

Основная работа по составлению корпуса была проделана [Барбарой Макгилливрэй](#), исследователем из института Алана Тьюринга, примерно в 2011 году, она же исправляла недочёты и обновляла ресурс в дальнейшем. Для лемматизации текстов был использован морфологический анализатор Дага Хауга. В качестве набора тегов используется [Lamap TAGset](#), разработанный Гельмутом Шмидом.

Цели при создании ресурса

Как утверждает создатель, создание корпуса - это заполнение ниши, образованной отсутствием достаточно удобного для анализа собрания латинских текстов. Из имеющихся на тот момент (2011 г.) лишь немногие проекты предоставляли возможность детального поиска, но даже они содержали в лучшем случае до 200000 токенов, остальные же в лучшем случае представляли из себя сырой текст, найти в котором можно было разве что словоформы.

Данный же корпус, при помощи современных инструментов NLP, предоставляет доступ к более чем 13 миллионам слов, содержит много полезных метаданных, размечает леммы и части речи, что и было основной целью проекта. Корпус нацелен прежде всего на исследование развития и эволюции латинского языка, так как охватывает латинские тексты, накопившиеся за 22 столетия, и прежде всего это касается лексики. Следовательно, предназначен он в основном для филологов и лингвистов, интересующихся диахронией. Более подробная информация описана в [статье, посвящённой ресурсу](#).

Благодаря уже имеющемуся функционалу, можно узнать, когда в употребление стали входить те или иные слова, когда стали реже употребляться те или иные части речи (скажем, герундив), какие есть коллокации у слов и для какого периода это характерно (например, **Deus** стал **Dominus** лишь с приходом христианства, а до этого в основном ассоциировался лишь с **templum**) и многое многое другое.

Языковой материал и состав текстов

Корпус был составлен на основе материалов из следующих источников:

1. [Intratext](#) - библиотека, содержащая множество книг на тему религии, науки, философии и прочего, содержит как классические латинские тексты, так и современные.
2. [Musisque Deoque](#) - архив, в котором собрана большая коллекция латинской поэзии вплоть до времён итальянского Ренессанса.
3. [LacusCurtius](#) - веб-сайт, посвящённый древнему Риму, на котором можно найти работы многих античных авторов

Все вышеперечисленные сайты предоставляют также разного рода метаинформацию, начиная от конкретного века и автора, заканчивая языком оригинала и метрикой.

Самые первые тексты датируются вторым веком до н.э., а самые поздние - концом XX века. Жанровое разнообразие также велико.

В качестве тренировочного материала для автоматической разметки были использованы данные с сайтов [Index Thomisticus Treebank](#) и [Latin Dependency](#)

[Treebank](#). Подробнее о том, на каких текстах базируется корпус можно узнать, перейдя по ссылкам выше, а также на сайте [самого корпуса](#).

Уровни разметки

Все тексты условно делятся на эры: классическую латынь (до II в. до н.э.), классическую латынь (I в. до н.э.), постклассическую латынь (I-VI вв), средневековую латынь (VI-XV вв.) и новую латынь (XV-XXI вв.). Дополнительно есть деление по конкретным векам, авторам, жанрам, названиям, сайтам-источникам и по языку оригинала.

Все тексты разбиваются на отдельные книги, если они есть, документы, строки (для поэзии), секции, абзацы и предложения.

Далее предложения делятся на токены, для каждого из которых можно посмотреть лемму.

Для морфологической разметки применяются автоматизированные методы. Введены отдельные POS теги для разных частей речи, нелатинских слова и знаков пунктуации, часть речи распознаются автоматически, что обеспечено методами машинного обучения для анализатора Дага Хауга.

На текущий момент это все доступные уровни разметки, хотя предполагается, что в будущем их станет больше. Во всяком случае планируется расширить морфологическую и добавить синтаксическую.

Возможности поиска

Ресурс позволяет искать отдельные слова, леммы, символы и части речи, есть поиск с использованием регулярных выражений, соответствия выдаются в формате KWIC. Всё это возможно в том числе с использованием CQL, причём доступен даже специальный конструктор для этого (рис. 1).

The screenshot displays the search interface for the LatinISE corpus. At the top, there is a navigation bar with a hamburger menu icon, the text "CONCORDANCE", a search input field containing "LatinISE corpus", a red "SUBSCRIBE" button, and a user profile icon. Below the navigation bar, there are three tabs: "BASIC", "ADVANCED", and "ABOUT". The "ADVANCED" tab is selected. On the left side, under "Query type", a dropdown menu is open, showing options: "simple", "lemma", "phrase", "word", "character", and "CQL". The "CQL" option is selected. In the center, there is a CQL query input field containing the query: `[lemma="book"][]{1,3}[tag="V.*"]`. Below the input field, there is an "Insert" button and a row of symbols: `[]`, `{}`, `<>`, `"`, `&`, `\`, `|`, `~`, and `#`. Below these symbols are two buttons: "TAGS" and "CQL BUILDER". At the bottom, there is a "Default attribute" dropdown menu set to "lemma". A "GO" button is located in the bottom right corner.

рис. 1
Интерфейс поиска

При необходимости, можно загрузить результаты поиска для работы оффлайн, правда, сначала всё равно потребуется выполнить поиск с Интернетом. К тому же скачивать можно лишь фиксированное количество токенов. Результаты сохраняются целиком, вместе со всеми текстовыми тегами.

Для коллокаций существует отдельный инструмент Word Sketch, который показывает самые вероятные варианты справа и слева от леммы. Он автоматически делит список коллокаций на части речи (рис. 2).

рис.2

WORD SKETCH DIFFERENCE

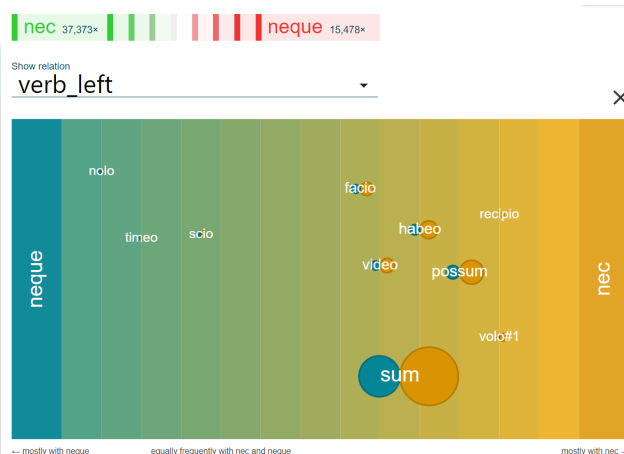


рис.3

То же можно сделать для двух слов сразу и сравнивать их коллокации одновременно. Можно визуализировать эти данные на диаграмме-области, каждый край которой показывает, с чем чаще ассоциируется слово (рис. 3).

Дополнительно для всех результатов поиска можно посмотреть на частотности и коллокации слов и словосочетаний, отфильтровать и отсортировать по наличию/отсутствию чего-либо, по словоформе, лемме, тегу и метаданным. Выбрать это можно на верхней панели, интерфейсы достаточно похожи друг на друга (рисунок ниже).

Для поиска синонимов существует инструмент Thesaurus, генерирующий наиболее похожие по семантике слова, его применение можно увидеть ниже.

Примеры запросов

Самое просто, что можно увидеть при помощи этого ресурса, это распределение частей речи между словами. Скажем, известно, что слово **cum** может быть как предлогом, так и союзом (в сослагательных предложениях). Корпус же позволяет узнать все возможные варианты, даже те, о которых сразу не догадаться, причём с примерами. Поскольку этот пример кажется наиболее интересным, стоит взглянуть именно на него поподробнее, пусть он и описан самим автором.

Simple search ?

cum

Text types ? ▾

SEARCH

Для начала осуществляем простой поиск по лемме.

Далее фильтруем по тегам и смотрим на возможные соответствия.

simple cum 88,849 (6.641.15 per million)

🔍 ⬇️ ≡ ⌂ 🔗 ✂️ ⚙️ ⚖️ 📄 ⚡ ⌨️ KWIC ▾ + ⓘ ☆

FREQUENCY

BASIC

ADVANCED

ABOUT

First word to the left

WORD FORMS

PART OF SPEECH

TAGS

LEMMAS

KWIC

WORD FORMS

PART OF SPEECH

TAGS

LEMMAS

First word to the right

WORD FORMS

PART OF SPEECH

TAGS

LEMMAS

More presets

TEXT TYPES

LINE DETAILS

☐ Details

Left context

KWIC

Right context

1

☐ poetry • uncert... exult oris artificis zonas et totas miscet amore. </s><s> talis Phoebeus erat,

cum

laetus caede draconis docta repercusso generavit carmina plectro: caelest

2

☐ poetry • uncert... ><s> est procul a nobis infelix gloriae Sullae trinaque tempestas, moriens

cum

Roma supremas desperavit et Martia vendidit arma. </s><s> nunc tellus in

3

☐ prose • Iulius ... acundiae retardavit et ab isto scribendi studio dubia trepidatione revocavit,

cum

fragilitas ingenii mei nihil se scire tale posse conciperet, quod dignum fore

4

☐ prose • Iulius ... e posse conciperet, quod dignum fore tuis auribus iudicaret. </s><s> Nam

cum

esses in Campaniae provinciae fascibus constitutus, cuius te administratio

5

☐ prose • Iulius ... nium enisus es fidis et religiosissimis amicitiae relevare fomentis. </s><s>

cum

itaque ad pristinum me statum solacis ac medelis tuis sanitas restituta rev

	Tag	↓ Frequency	Per million tokens		
1	<input type="checkbox"/> PRE	59,629	4,457.06	<div></div>	...
2	<input type="checkbox"/> C	29,215	2,183.72	<div></div>	...
3	<input type="checkbox"/> N	3	0.22	<div></div>	...
4	<input type="checkbox"/> ADJ	2	0.15	<div></div>	...

Результаты показывают, что слово **cum** в подавляющем числе случаев является предлогом, иногда союзом, и крайне редко существительным или наречием.

Если необходимо, можно посмотреть на случаи употребления, например, в качестве существительного.

simple **cum** 3 > filter [tag="N"] 3 (0.22 per million) X

Left context KWIC Right context

1	<input type="checkbox"/>	prose • Iulius ...	</s><s> Illud etiam diligenti debemus ratione colligere, et cum benivola et cum malivola dominium fuerit geniturae consecuta, an in oportunis geniturae loc
2	<input type="checkbox"/>	prose • incerta...	, talem ad praefectum urbis super morte Cari epistulam dedit: Inter cetera" Cum ," inquit," Carus, princeps noster vere carus, aegrotaret, tanti turbinis subit
3	<input type="checkbox"/>	prose • Isidoru...	tio est. </s><s> Nam' huiusce' per C,' cuiusque' per Q scribimus. </s><s> ' Cum ' autem praepositio per C scribenda est; si autem adverbium fuerit, per Q. I

Pyrrhum'. </s><s> C et G litterae quandam cognationem habent. </s><s> Nam dum dicimus' centum', ei' 'trecentos', postea dicimus' quadringentos', G ponentes pro C. C et Q similiter cognatio est. </s><s> Nam' huiusce' per C,' cuiusque' per Q scribimus. </s><s> ' **Cum** ' autem praepositio per C scribenda est; si autem adverbium fuerit, per Q. Dicimus enim' quum lego'. </s><s> ' Deus' per E solam: ' daemon' per AE diphthonga est notandus. </s><s> ' Equus', quod est animal, per E solam scribendum. </s>

Здесь это значит, что слово **cum**-существительное представляет собой просто название слова. Ничего совсем необычного мы не узнали, но всё равно полезно помнить и о таких случаях.

Другой пример: почему-то при поиске некоторых слов по лемме у них встречаются странные, непривычные основы, пример ниже - частотный список словоформ слова **caelum**.

	Word	↓ Frequency	Per million tokens	
1	<input type="checkbox"/> caelo	2,223	166.16	...
2	<input type="checkbox"/> caelum	2,040	152.48	...
3	<input type="checkbox"/> caeli	1,834	137.09	...
4	<input type="checkbox"/> caelis	306	22.87	...
5	<input type="checkbox"/> caelorum	224	16.74	...
6	<input type="checkbox"/> Caelum	75	5.61	...
7	<input type="checkbox"/> Caelo	34	2.54	...
8	<input type="checkbox"/> celis	16	1.20	...

У последнего из них на месте **ae** стоит почему-то **e**. Если посмотреть на слово в контексте, оно будет значить то же, что и **caelis**. Разгадка кроется в датах, если посмотреть, когда это слово употреблялось, окажется, что почти всегда это средневековая латынь.

	Century	↓ Frequency	Relative % ?	
1	<input type="checkbox"/> cent. 14 A. D.	5	2,534.2	...
2	<input type="checkbox"/> cent. 12-13 A. D.	3	1,787.7	...
3	<input type="checkbox"/> cent. 13 A. D.	2	957.4	...
4	<input type="checkbox"/> cent. 11 A. D.	2	1,395.8	...
5	<input type="checkbox"/> cent. 5 A. D.	1	549.8	...
6	<input type="checkbox"/> cent. 4 A. D.	1	150.9	...
7	<input type="checkbox"/> cent. 13-14 A. D.	1	5,964.7	...
8	<input type="checkbox"/> cent. 10 A. D.	1	1,625.1	...

И действительно, судя по всему, это отражает традиционное латинское произношение. К сожалению, здесь не показываются все возможные падежи, поэтому для пущей уверенности поиск придётся немного расширить, возможно, это как раз небольшая

недоработка корпуса. Ниже распределение для слова **celum**.

	Century	↓ Frequency	Relative % ?	
1	<input type="checkbox"/> cent. 14 A. D.	27	2,575.9	...
2	<input type="checkbox"/> cent. 12-13 A. D.	25	2,804.2	...
3	<input type="checkbox"/> cent. 13-14 A. D.	16	17,964.3	...
4	<input type="checkbox"/> cent. 12 A. D.	7	138.9	...
5	<input type="checkbox"/> cent. 16 A. D.	2	90.3	...
6	<input type="checkbox"/> cent. 15 A. D.	2	183.7	...
7	<input type="checkbox"/> cent. 10 A. D.	2	611.8	...
8	<input type="checkbox"/> cent. 5-7 A. D.	1	60.9	...
9	<input type="checkbox"/> cent. 17 A. D.	1	45.5	...
10	<input type="checkbox"/> cent. 13 A. D.	1	90.1	...

И наконец, самое интересное - составим список синонимов для слова **celum** с помощью Thesaurus, чтобы убедиться наверняка.

	Word	Frequency ?
1	Fronesis	29 ...
2	celsitudinem	22 ...
3	coelum	204 ...
4	indeque	87 ...
5	caelus	240 ...

Как можно заметить, самые частотные варианты это как раз **coelum** и **caelus**. Оба эти слова - ещё одни варианты слова **caelum** и они, опять же, начинают употребляться ближе к Средневековью, в чём предлагается удостовериться самостоятельно, если есть желание. Причём выходит, что это небо как раз в религиозном представлении, у слова **caelum** список синонимов несколько шире. Таким образом, мы увидели, как менялось слово в мёртвом, казалось бы, языке.

Дополнительные ресурсы

Основная статья, использованная для написания письменной части:

McGillivray, B. and Kilgariff, A. (2013). *Tools for historical corpus research, and a corpus of Latin*. - URL:

https://www.sketchengine.eu/wp-content/uploads/2015/05/Latin_historical_corpus_2013.pdf

В более сжатом виде то же можно найти на сайте корпуса:

<https://www.sketchengine.eu/latinise-corpus/>

Отдельных исследований, посвящённых корпусу нет, фактически вся информация ограничена той, что уже была приведена выше.

Однако есть информация касательно SketchEngine в целом, что может упростить работу с ресурсом:

- Видео-инструкция:
https://www.youtube.com/watch?v=f4eszLB47Qk&feature=emb_title
- Руководство пользователя:
<https://www.sketchengine.eu/guide/>

В самом дизайне сайта также достаточно много подсказок по тому, как работает тот или иной элемент корпуса, отсылающих на соответствующие статьи.