

# Using NLP to Create Corpus-based Vocabulary Exercises in Latin Classes

Machina Callida

Andrea Beyer    Konstantin Schulz

Humboldt Universität zu Berlin

March 3rd, 2020

Funded by



Deutsche  
Forschungsgemeinschaft

German Research Foundation



# Outline

- 1 Motivation
- 2 Enhanced Latin Vocabulary Learning
- 3 Outcome of Corpus-based Vocabulary Learning

# Scope

- Language: Latin
- Institution: high school
- Location: Berlin
- Target audience: teachers and intermediate learners
- Context: vocabulary training
- Problem so far:
  - ▶ Infrequent training (homework)
  - ▶ Inefficient methods (rote and cross-lingual learning, single words without context)
  - ▶ Decoupled from syntax and morphology

# Solution: Vocabulary Software

- The software is accessible at all times, on all platforms.
- It offers authentic contexts, e.g. phrases or small texts.
- It also offers ready-made exercises for instant usage.
- It supports the creating of personalized exercises.

The screenshot displays the Machina Callida web application interface. At the top, the logo and name 'Machina Callida' are on the left, and 'English' with a menu icon is on the right. Below the header, a main heading reads 'Context matters: Learn to use Latin words smartly!'. A paragraph of text follows, explaining the software's focus on authentic contexts and citing John Rupert Firth's 1957 motto: 'No exercise without a reference to the context of the word, as the English linguist John Rupert Firth wrote in 1957: "You shall know a word by the company it keeps."'.

The interface is divided into four main sections, each with a 'CONTINUE' button:

- Create exercise:** Includes options for 'Selection of text', 'Text complexity', 'Compare vocabulary', and 'Exercise parameters'.
- Exercise Repository:** Lists 'Exercises created' with categories: 'Cloze', 'Mark Words', and 'Matching'.
- Vocabulary unit:** Lists 'Text work', 'Exercises', 'Final test', and 'Evaluation'.
- Documentation:** Lists 'About the project', 'Software doc', 'Exercises doc', and 'Vocabulary unit doc'.

At the bottom of the page, there are logos for 'Inprint' and 'Refresh corpora'.

# Authentic and context-based exercises

## Most recent settings:

Hieronymus (PROIEL), Vulgata (Novum Testamentum), 1.6-1.9

High-quality texts only (uncheck for all authors)

🔍

## Authors

[C. Iulius Caesar \(PROIEL\)](#)

[Hieronymus \(PROIEL\)](#)

[M. Tullius Cicero \(PROIEL\)](#)

[Palladius \(PROIEL\)](#)

[Peregrinatio Aetheriae \(PROIEL\)](#)

- Exercises to be selected from more than 100 Latin authors.
- The corpus is of varying quality depending on the annotations (manual vs. automatic).

Mark Words: Mark words that fit at least one of the given descriptions! [Conjunction](12)

Quamquam <sup>(+)</sup>  Marco fili annum iam audientem Cratippum id que Athenis abundare oportet praeceptis institutis que philosophia <sup>(+)</sup>  propter summam et doctoris auctoritatem et urbis quorum alter te scientia augere potest altera exemplis tamen <sup>(-)</sup>  ut <sup>(-)</sup>  esse ad meam utilitatem semper cum Graecis Latina <sup>(+)</sup>  coniungi <sup>(-)</sup>  ne <sup>(-)</sup>  que id in philosophia solum sed etiam in dicendi exercitatione feci idem tibi censeo faciendum <sup>(+)</sup>  par sis in utriusque orationis facultate. Quam quidem ad rem nos ut videmur magnum adumentum hominibus nostris <sup>(+)</sup>  non modo Graecorum litterarum rudes sed etiam docti aliquantum se arbitrentur adeptos et ad dicendum et ad iudicandum.

Score: 1 of 12.

1/12

⌂ Retry

🔑 Solution

DOCX

PDF

[Report an error](#)

BACK

CHANGE TEXT PASSAGE

🔗 SHARE

- The exercises work intra-lingually, e.g. cloze.
- The exercises focus on form, function or meaning, e.g. mark words.

# Preparing the exercises

## Selected Text

### Highlight unknown vocabulary

Gallia est omnis divisa in partes tres quarum unam incolunt Belgae aliam Aquitani tertiam qui ipsorum lingua Celtae nostra Galli appellantur. Hi omnes lingua institutis legibus inter se differunt. Gallos ab Aquitanis Garumna flumen a Belgis Matrona et Sequana dividit. Horum omnium fortissimi sunt Belgae propterea quod a cultu atque humanitate provinciae longissime absunt minime que ad eos mercatores saepe commeant atque ea quae ad effeminandos animos pertinent important proximi que sunt Germanis qui trans Rhenum incolunt quibus cum continenter bellum gerunt. Qua de causa Helvetii quoque reliquos Gallos virtute praecedunt quod ferè cotidianis proeliis cum Germanis contendunt cum aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt.

### Text complexity

Overall complexity: 37.03  
 Word count: 116  
 Sentence count: 5  
 Words per sentence (Ø): 23.2  
 Word length (Ø): 5.59  
 Number of different word forms: 92  
 Number of different parts of speech:  
 13  
 Lexical density: 0.63  
 Punctuation mark count: 5  
 Main clause count: 5  
 Subclause count: 7  
 Infinitive count: 0  
 Participle count: 1  
 Gerund count: 0  
 Number of Ablativi Absoluti: 0

- You can compare the text to your personal vocabulary.
- You can assess the linguistic complexity of the chosen text.
- You can set different parameters: interaction type, linguistic phenomena, instructions.

# Working with the exercises

**Exclude unknown words**

Cloze: Assign the words from the pool to the correct gaps!

Quamquam te Marce fili annum iam audientem Cratippum id que exemplis ✘ abundare oportet praeceptis institutis Athenis  
 que philosophiae propter summam et doctoris auctoritatem et urbis quorum alter te scientia augere potest altera  
  tamen ut ipse ad meam utilitatem semper cum Graecis ✔ Latina coniunxi ne que id in philosophia  
 solum sed etiam in dicendi exercitatione feci idem tibi censeo faciendum ut par sis in utriusque orationis facultate.  
 Quam quidem ad rem nos ut videmur magnum attulimus adiumentum hominibus nostris ut non modo Graecarum  
 litterarum rudes sed etiam docti aliquantum se arbitrentur adeptos et ad dicendum et ad iudicandum.

**Score: 1 of 3.**

★ 1/3

Solution

Retry

[DOCX](#)
[PDF](#)
[XML](#)
[Report an error](#)

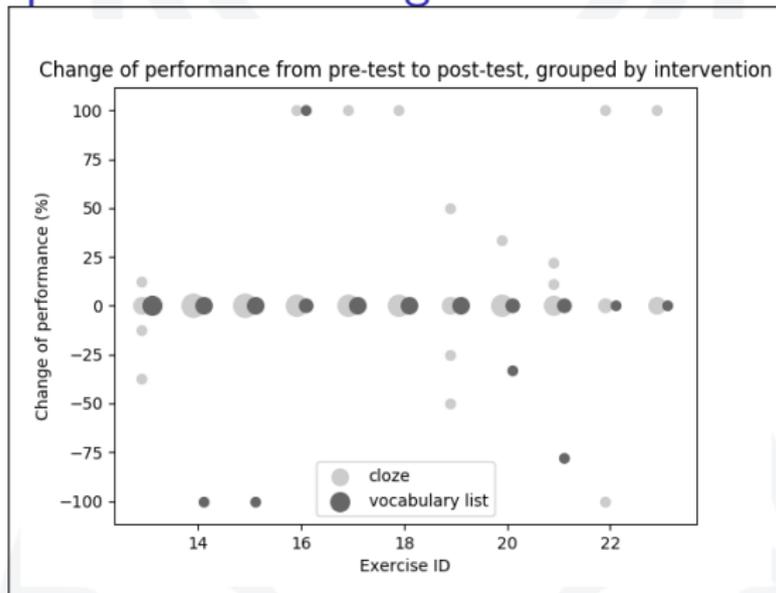
BACK

CHANGE TEXT PASSAGE

↪ SHARE

- You get an automatic feedback (and scores).
- You can share a link or use exercises stored in the repository.

## Study on corpus-based learning: Cloze vs. word list



Comparison of students' performance before and after interventions. Larger markers indicate multiple students with the same percental change. We may conclude carefully that solving cloze exercises is correlated with improvements in vocabulary competence (more light grey dots above the baseline), as opposed to traditional rote learning.

## Conclusions

- ① Performance improvements are not restricted to just matters of word choice or the translation of single words. Instead, they even relate to more complex tasks like reading comprehension or translation of multiword expressions.
- ② Therefore, lexical competence is to be understood in a broad sense, i.e. including certain aspects of syntax and morphology.
- ③ Finally, corpus-based vocabulary training seems to be beneficial for a learner's lexical competence.

Try our software (Machina Callida)  
<https://korpling.org/mc> !

