FROM TEXTS TO **EXERCISES (SEMI)AUTOMATICALLY**

Neven Jovanović (neven.jovanovic@ffzg.hr) Berlin, October 4-5, 2017

This page: croala.ffzg.unizg.hr/callidus-automatic-exercises/

Repository: bitbucket.org/nevenjovanovic/discipulus

THE PLAN

Preliminary thoughts The components Case study: exercises for a reading corpus What is missing?

PRELIMINARY

THE ROLE: SUPPLEMENT AND PREPARATION

WHICH TYPES OF EXERCISES?

THE FORCE OF IMAGES, Sound, Actions

THE PROTOTYPE ISSUE

THE COMPONENTS

1. THE TEXTS

"Here are the texts I want, or need, to read — can you help me read them?"

2. GRAMMATICAL ANNOTATIONS

LEMLAT

3. VOCABULARIES

Dickinson College Core

(... auf Deutsch?)

4. MANPULATION

BaseX and XQuery

CLTK and Python

5. PUBLICATION

Moodle: more control to the **teacher** H5P: works with Moodle, excellent design Anki: more control to the **learner** Both Moodle and Anki are accessible from **smartphones** as well!

FROM TEXTS TO A READING CORPUS

MANY SMALL STEPS...

THE TEXTS, **TOKENIZED** INTO **CHAPTERS SENTENCES PHRASES** WORDS

FREQUENCIES FOR THE WHOLE CORPUS FOR TEXTS **FOR CHAPTERS** (EVEN FOR SENTENCES?)

LEMMATA FOR TOKENS

connect lemmata with tokens!

FREQUENCIES FOR LEMMATA

concentrate on what is frequent – or on what is rare!

ALIGNMENTS OF TRANSLATIONS... WITH LEMMATA WITH OCCURRENCES WITH PHRASES WITH SENTENCES

DISTRACTORS FOR MULTIPLE CHOICE ETC. Sound...? Video...?

(Cf. the Memrise "Meet the natives" feature.)

CASE STUDY: READING LIST FOR THE "TRANSLATION FROM LATIN" COURSE AT THE UNIVERSITY OF ZAGREB

Seneca, Letters to Lucilius 1 Terence, Adelphoe Horace, Odes 1 Tibullus 1

COMPILE THE CORPUS

GET THE TEXTS FROM THE PERSEUS DL

All texts are available there, but the task was not quite trivial, because we needed just some segments (some books) of these works, and also — as it turns out — because the texts contain critical notes and, in two places, funny division of words (Horatian metres!).

CREATE WORD LISTS AND FREQUENCIES

The corpus contains 24,318 words.

Seneca: 6272 Terence: 8874 Horace: 3967 Tibullus: 5211

Feed the word list to LEMLAT for lemmatization analysis. The results:

Number of word forms: 8548 (different forms in 24,318 words) Number of forms unknown to the program: 103 Number of forms analysed: 8445

ANNOTATE LEMLAT ANALYSES For single or multiple candidates

Distinguish forms with multiple candidate lemmata (multilem) from the forms with unambiguous lemmata (both derived and unique).

Some LEMLAT lemmata are homonyms, or even identical, though they have different LEMLAT id numbers; if we isolate such cases, the number of annotated forms can be enlarged further.

AN OPPORTUNITY TO ENGAGE STUDENTS?

TOKENIZE THE TEXTS INTO Sentences and words

For sentences, CLTK has the best tool; we want to use the tool through a pair of Python scripts:

one which reads all text files, the other which tokenizes the text in them into sentences.

(We also need to prepare text-only versions of our segments.)

Then we reconstruct from JSON files (output by the CLTK tool) the text documents, with all their chapters, scenes, letters, and poems, but now also with sentences below these levels (and tokenized into words and punctuation as well). For that, we have two XQuery scripts: one to annotate punctuation, the other to tokenize the remaining text nodes.

ANNOTATE WORD FORMS IN CORPUS WITH Pointers to lemlat lemmata

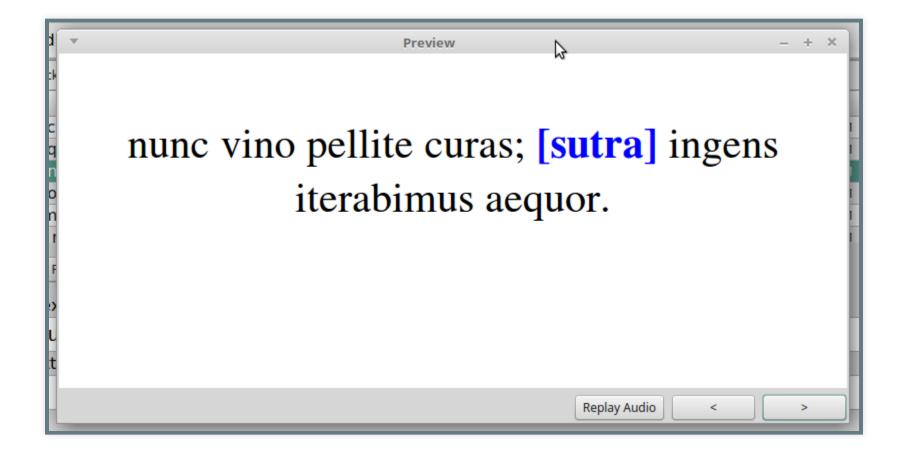
<w lemma="homo"
lemmaRef="lemlat:h234">hominem</w>
Automatically, a little over 48% of words in the corpus is
lemmatized.

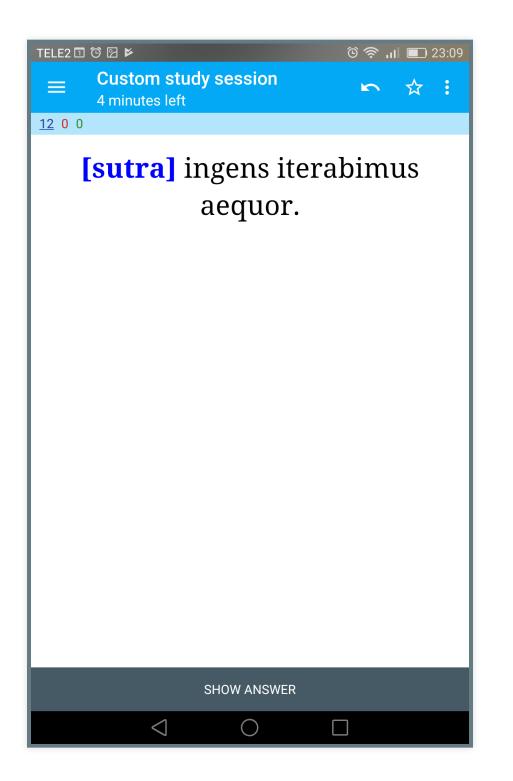
CREATE EXERCISES

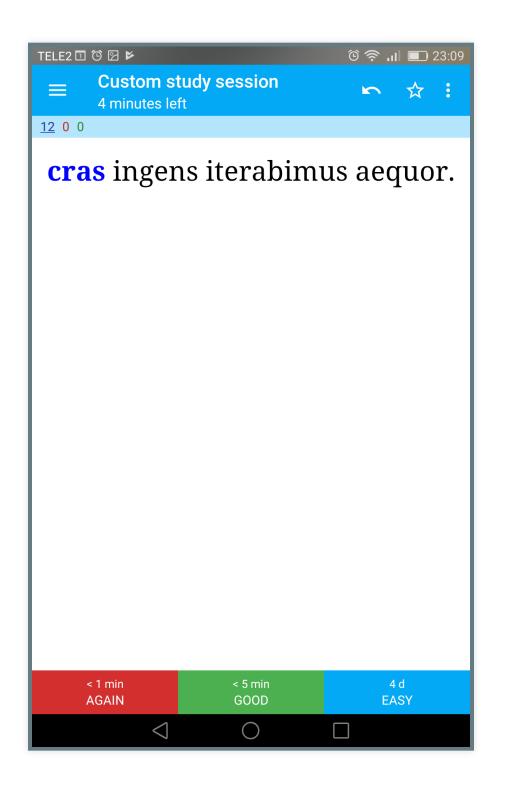


Practise vocabulary and sentences

{{cl::crastinum::sutrašnji dan}} si adiecerit deus, laeti recipiamus. cras
{{cl::crastinum::sutrašnji dan}} sine sollicitudine expectat. cras
{{cl::cras::sutra}} ingens iterabimus aequor. cras
nunc vino pellite curas; {{cl::cras::sutra}} ingens iterabimus aequor. cras







INTERESTING EXERCISE TYPES

Latin word / Croatian meaning and reverse (to introduce a word) Sentence with a Croatian translation, and reverse Cloze card with the word to be supplied in Croatian (in its vocabulary form) Smaller, meaningful phrases from sentences with translation and clozes Croatian words in vocabulary form, ordered as in the sentence (produce the correct sentence in Latin!)

MOODLE – DATABASE Activity

Add translations

Omega (= Moodle instance of the Faculty of Humanities and Social Sciences, University of Zagreb)

MOODLE - QUESTION BANK

Mass import of automatically generated Q&A cards

Omega

WHAT IS MISSING?