Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

LEA: Linguistic Exercises with Annotation Tools

Fabian Barteld
fabian.barteld@uni-hamburg.de

Johanna Flick
johanna.flick@uni-duesseldorf.de

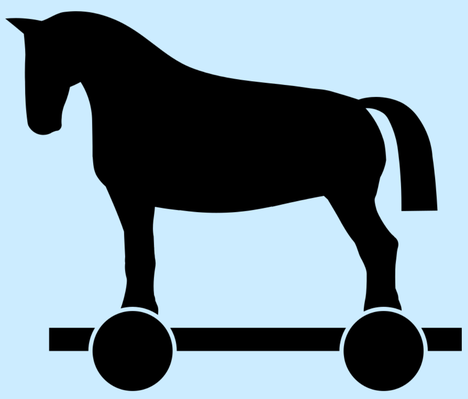
What is LEA?

"Working empirically is a complex task and acquiring the necessary technical expertise in order to use the tools can lead to frustration" (Bubenhofer, 2011, p. 148)

LEA (**L**inguistic **E**xercerises with **A**nnotation Tools) is a new didactic concept helping students to become familiar with corpus linguistic methods and annotation tools.

Main idea

Linguistic exercises like PoS categorization or syntactic analysis are being solved with annotation tools. The didactic concept can be best explained in terms of the "Trojan Horse metaphor" except that there is something good coming out from the horse's inside, i.e. the computational know how: While doing exercises students learn about corpus linguistic methods and ways of creating sustainable annotated data, since they are basically solving an annotation task.



Advantages for students

- Step by step introduction to annotation tools and guidelines
- Media change and new visualizations can help to understand the linguistic concepts

Advantages for teachers

- Tools to automatically analyze and evaluate students' answers
- Ready to use in linguistic classes
- Easy integration of own exercises

The part-of-speech exercise

Every student of linguistics is being confronted with PoS exercises at some point of her academical studies, usually in the first year. A classical task would be assigning categories like noun, verb or preposition to specific word forms, e.g. the tokens in a sentence.

Die	Katze	sitzt	unter	dem	Tisch
Det.	Noun	Verb	Prep.	Det.	Noun

Task
Use the simplified version of STTS for the parts of speech and add the tags to the sentence in a tab-separated-values file using a spreadsheet application.

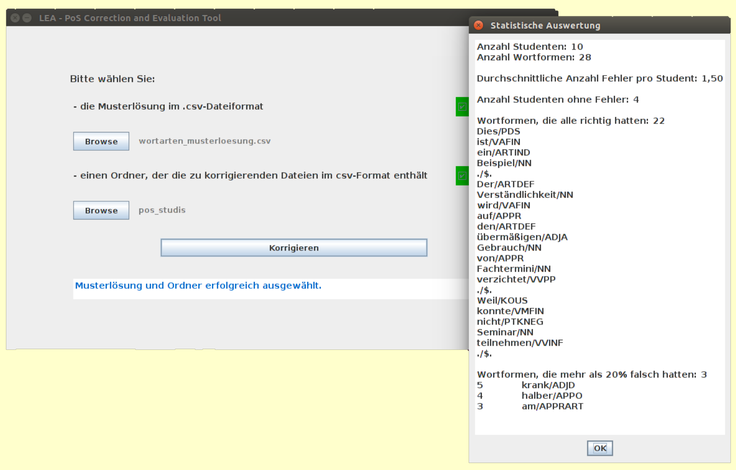
Token	STTS
Die	ARTDEF
Katze	NN
sitzt	VVFIN
unter	AP
dem	ARTDEF
Tisch	NN

Learning goals

(1) Various aspects connected to csv or similar files:
- Opening them with a spreadsheet application,
- handling different kinds of delimiters like tab or comma,
- choosing the right encoding when opening text files which contain non-ASCII characters like German umlauts.

(2) STTS and annotation guidelines

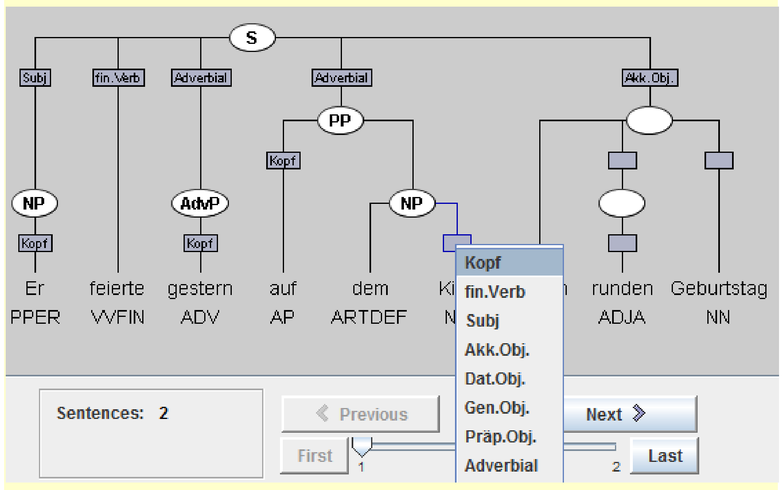
Automatic correction



The syntax exercise

Students of German linguistics need to practice how to (1) identify phrases, (2) assign the right category to them and (3) assign syntactic functions to them.

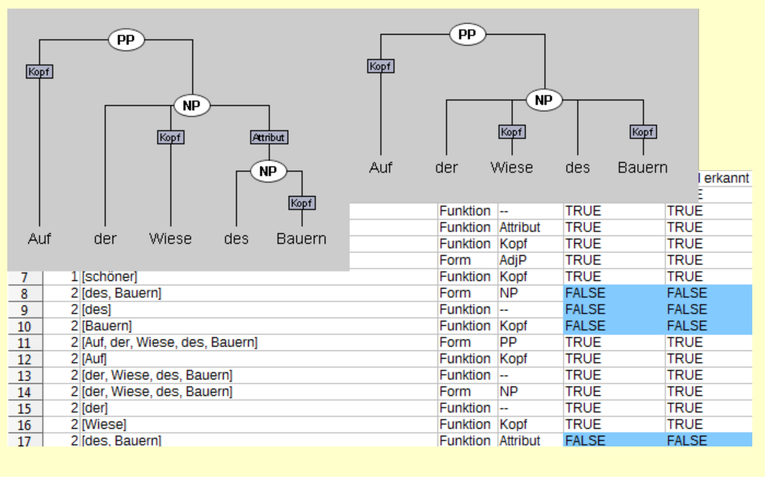
Task
Use Synpathy to mark the phrases of a sentence, assign the phrase types as node labels and syntactic functions as edge labels.



Learning goals

(1) Becoming familiar with trees (nodes & edges) for syntactic annotation
(2) First contact with xml files
(3) Annotation scheme (similar to the Negra/Tiger-Annotation scheme cf. Brants et al., 2002)

Automatic correction



Present and future of LEA

The two presented LEA exercise packages are used in introductions to German linguistics at the Universities of Hamburg and Düsseldorf. Based on the experiences from these classes we improved the packages. We included, for instance, descriptions on how to download the xml files for the syntax exercise which proved to be problematic.

For the future, we plan to add further exercises which address new topics and introduce new annotation tools, e.g. annotation of semantic roles with SALTO (Burchardt et al., 2006), spoken language with EXMARaLDA (Schmidt and Wörner, 2009) or phonetic analysis with Praat (Boersma and Weenink, 2013).

Find LEA online

https://korpuslab.github.io/lea

References

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius and George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, pages 24–41.

Paul Boersma and David Weenink. 2013. Praat: doing phonetics by computer. http://www.praat.org/.

Noah Bubenhofer. 2011. Korpuslinguistik in der linguistischen Lehre: Erfolge und Misserfolge. Journal for Language Technology and Computational Linguistics (JLCL), 26(1):141–156.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. SALTO – a versatile multi-level annotation tool. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006).

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Universities of Stuttgart und Tübingen.

Thomas Schmidt and Kai Wörner. 2009. EXMARaLDA – creating, analysing and sharing spoken language corpora for pragmatic research. Pragmatics, 19(4):565–582.

Synpathy = https://tla.mpi.nl/tools/tla-tools/older-tools/synpathy