Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

LEUPHANA
UNIVERSITÄT LÜNEBURG

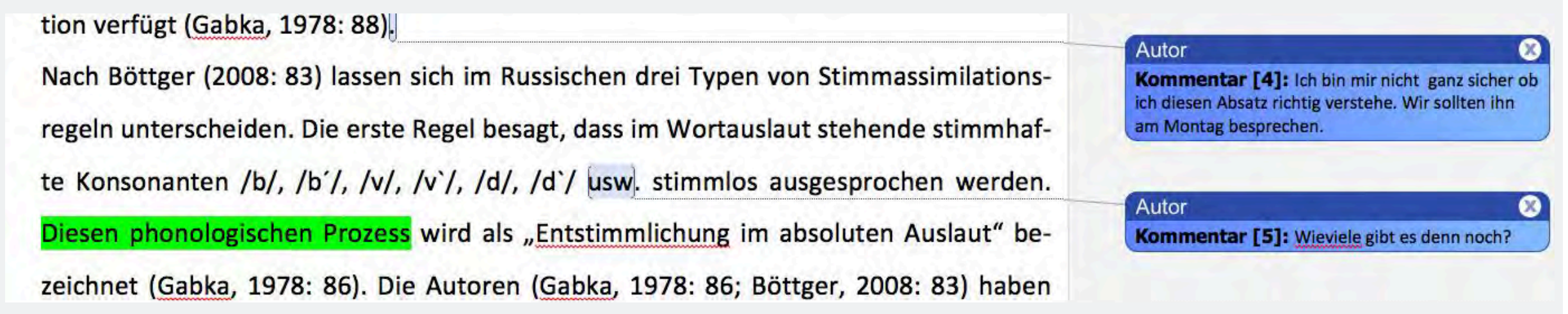
KoLaS – Commented Learner Corpus of Academic German

WHAT IS KOLAS?

KoLaS is short for 'Kommentiertes Lernendenkorpus akademisches Schreiben' ('Commented Learner Corpus of Academic German'). It is a corpus of German academic texts written by students of the University of Hamburg. The texts are highly authentic in that they were written for assessments during their studies.

Aims of KoLaS
Until now, there are no learner data for academic German that are publicly available. With the release of KoLaS, we wish to fill this gap and enable more verifiable research on academic learner German, text revision processes and commenting behaviour of peer tutors.

More information: https://www.korpuslab.uni-hamburg.de/kolas



tion verfügt (Gabka, 1978: 88).

Nach Böttger (2008: 83) lassen sich im Russischen drei Typen von Stimmassimilationsregeln unterscheiden. Die erste Regel besagt, dass im Wortauslaut stehende stimmhafte Konsonanten /b/, /b'/, /v/, /v'/, /d/, /d'/ usw. stimmlos ausgesprochen werden. Diesen phonologischen Prozess wird als „Entstimmlichung im absoluten Auslaut" bezeichnet (Gabka, 1978: 86). Die Autoren (Gabka, 1978: 86; Böttger, 2008: 83) haben

Autor
Kommentar [4]: Ich bin mir nicht ganz sicher ob ich diesen Absatz richtig verstehe. Wir sollten ihn am Montag besprechen.

Autor
Kommentar [5]: Wieviele gibt es denn noch?

Figure 1: Example file from the corpus with comments by a peer tutor

Ko LaS

Kommentiertes Lernendenkorpus akademisches Schreiben

('Commented Learner Corpus of Academic German')

WHAT KIND OF DATA ARE AVAILABLE?

The corpus comprises the texts themselves, some of them with comments by peer tutors, metadata about the writers and, if applicable, metadata about the consultation related to the text.

DATA COLLECTION

Context of data collection
At the 'Schreibwerkstatt Mehrsprachigkeit' (*writing centre multilingualism*), a writing center at the University of Hamburg (Knorr/Neumann 2014), students could get feedback on the texts they have to write during their studies. In return, the writing center is allowed to carry out analyses of the texts received this way and make the available for the academic public. Consequently, the corpus is not compiled according to a specific research question, but is rather what has been called an 'opportunistic corpus' (e.g. Teubert/Čermáková 2004:120). Depending on the research question, compilation of a subcorpus might be advisable.

Schreibwerkstatt Mehrsprachigkeit

Data processing
Version 2.0 of the corpus comprises texts from 2011 until 2016. All texts were made anonymous, including information about the writers, the lecturers and the peer tutors. In order to make metadata about the texts available, a metadata file for the tool Corpus-Manager (Coma) was created. Coma can be downloaded for free.

More information on Coma:
http://www.exmaralda.org/tool/corpus-manager-coma/

Texts and comments

- 854 texts in total
- written between September 2011 and December 2016
- by 112 different students
- text types: mainly seminar papers, bachelor and master thesis, internship reports, lecture reports
- more than half of the texts are commented on by our peer tutors (see figure 1 above)
- occasionally: several versions of one texts

Metadata about the writers

- field of study: predominantly students from educational science and the humanities
- semester: between 1 and 30, 6 on average
- languages learned at home, at school and languages their parents speak: answers include German, Russian, Bassa, Chinese, English, Persian, French, Greek, Hindi, Indonesian, Japanese, Kandahari, Kazakh, Catalan, Kyrgyz, Croatian, Kurdish, Multani, Slovak, Spanish, Sundanese, Czech, Turkish, Twi, Ukrainian, Yuruba
- per student the corpus includes 7 texts on average

Metadata about the consultations

- date
- duration
- persons involved (usually one student and one peer tutor)
- topics discussed (e. g. grammar, text structure, how to find a topic…)
- section of text discussed (e. g. introduction, literature review, discussion…)
- additional notes by the peer tutors
- occasionally: audio records and transcripts of consultations

RESEARCH & TEACHING WITH KOLAS

Research
So far, KoLaS has been used for the analysis **of learner academic language**, often in comparison to professional academic language (Andresen 2016 and several student papers). In a cooperation with the pedagogical university of Bern the data were used to validate an annotation scheme for **analysing commenting behaviour** (e.g. Beyer 2016, von Gunten 2015). Further possible research question concern text development, influence of culture and other languages and analysis of consultations (also in relation to written text comments).

Teaching
KoLaS is beeing used at the 'Schreibwerkstatt Mehrsprachigkeit' for our **peer tutor training** and in **introductions to academic writing**. It provides us with authentic examples and lets students explore issues in academic writing themselves ('serendipitous exploration', Aston 1998). At the same time, it introduces students to basic research methods.

FUTURE PLANS AND NEEDS

Increase user-friendliness
Currently, the data are available as Word files only. For the future, we wish to make the data more easily accessible by converting it into software-independent formats and enabling online queries (e. g. via ANNIS, Krause/Zeldes 2014). However, so far there is **no suitable infrastructure for commented texts** and possible annotations (to the texts and the comments, respectively).

Corpus extension
Thanks to a cooperation with several lecturers in educational sciences at the University of Hamburg in summer 2016, we will be able to create a **new subsection of about 300 essays** written by 100 students at three points of time during the semester.

OBTAINING KOLAS

KoLaS is **publicly available** and we wish to encourage other researchers to use the corpus for their own research or training purposes. The corpus is hosted by and can be obtained via the HZSK ('Hamburger Zentrum für Sprachkorpora', https://corpora.uni-hamburg.de/drupal/).
For more information on the corpus see https://www.korpuslab.uni-hamburg.de/kolas.

Melanie Andresen
Universität Hamburg
Institut für Germanistik
Melanie.Andresen@uni-hamburg.de

Dr. Dagmar Knorr
Leuphana Universität Lüneburg
Schreibzentrum / Writing Center
Dagmar.Knorr@leuphana.de

References

Andresen, Melanie. 2016. Im Theorie-Teil der Arbeit werden wir über Mehrsprachigkeit diskutieren – Sprechhandlungsverben in Wissenschafts- und Pressesprache. Zeitschrift für angewandte Linguistik 64(1). 47–66. doi:10.1515/zfal-2016-0001.

Aston, Guy. 1998. Learning English with the British National Corpus. In M. P. Battaner & C. López (eds.). Barcelona. http://www.sslmit.unibo.it/~guy/barc.htm.

Beyer, Anke. 2016. *InliAnTe: Instrument für die linguistische Analyse von Textkommentierungen* (Arbeitspapiere Projekt „Texte kommentieren", 1). Bern: PHBern.

von Gunten, Anne. 2015. Wie und wozu angehende Lehrpersonen Texte kommentieren. Forschungsdesign und Analyse-Instrumente eines Projektes. Zeitschrift Schreiben. http://www.zeitschrift-schreiben.eu.

Knorr, Dagmar & Ursula Neumann. 2014. Die Schreibwerkstatt Mehrsprachigkeit – (Lehramts-)Studierende mit Migrationshintergrund der Universität Hamburg schreiben. In Dagmar Knorr & Ursula Neumann (eds.), Mehrsprachige Lehramtsstudierende schreiben: Schreibwerkstätten an deutschen Hochschulen, 119–144. (FörMig Edition 10). Münster: Waxmann.

Krause, Thomas & Amir Zeldes. 2014. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities* 31(1). 118–139. doi:10.1093/llc/fqu057.

Teubert, Wolfgang & Anna Čermáková. 2004. Directions in corpus linguistics. In M. A. K. Halliday, Wolfgang Teubert, Colin Yallop & Anna Cermáková (eds.), *Lexicology and Corpus Linguistics. An Introduction*, 113–165. (Open Linguistics Series). London, New York: Continuum.

www.korpuslab.uni-hamburg.de/kolas