Corpus-supported academic writing: how can technology help?

Madalina Chitez[1], Christian Rapp[2], and Otto Kruse[3]

Abstract. Phraseology has long been used in L2 teaching of academic writing, and corpus linguistics has played a major role in the compilation and assessment of academic phrases. However, there are only a few interactive academic writing tools in which corpus methodology is implemented in a real-time design to support formulation processes. In this paper, we describe several corpus-related methods that we have developed and implemented as part of an interactive thesis-writing tool, *Thesis Writer*, designed and constructed jointly by the Language Competence Centre and the Center for Innovative Teaching and Learning of the Zurich University of Applied Sciences in Switzerland. *Thesis Writer* (TW) hosts several linguistic-support tools and is designed in its first pilot version to support thesis writing in economics with the help of two self-compiled corpora in English and German. Students can access the corpora directly via the IMS Open Corpus Workbench or via a pre-selected collection of central rhetorical elements through the phrase book. Several search options and tutorials have been tested and included into the TW platform: the corpus simple search tool, the corpus syntactic search tool, and the academic phrasebook. In the case of the latter, a new methodology led to the identification of lists of phrases distributed in research-cycle sections of the thesis.

Keywords: academic writing, corpus linguistics, language instruction, thesis writing, educational platform, collocation, academic phrases, academic phrasebank, L2 German, L2 English.

1. Zurich University of Applied Sciences, Switzerland; madalina.chitez@zhaw.ch
2. Zurich University of Applied Sciences, Switzerland; christian.rapp@zhaw.ch
3. Zurich University of Applied Sciences, Switzerland; otto.kruse@zhaw.ch

How to cite this article: Chitez, M., Rapp, C., & Kruse, O. (2015). Corpus-supported academic writing: how can technology help? In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 125-132). Dublin: Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2015.000321

1. Introduction

Since the 1970s, when the teaching of writing began to shift from a text-oriented to a process-centered approach, writing instruction has largely abstained from a direct teaching of language. From this time on, the interest of teachers and researchers has focused on what the writers do and think rather than on the linguistic or textual means they use. Recently, several experts have demanded a reconsideration of the role of language in writing (e.g. Feilke 2010, 2012, 2014; Hyland, 2000; Myhill & Fischer, 2010; Steinhoff, 2007) and started to develop theoretical and educational models. However, it is yet to be explored what a linguistically informed writing process might look like and how the formulation process can be supported by knowledge about language. As Feilke (2010, 2012) suggested, a plausible assumption is that writing relies on a high number of routinized textual procedures, which serve rhetorical and structural functions in the construction of meaning.

Corpus linguistics provides several effective approaches at the interface of research in learner language and academic writing, which can be used to identify such routines as phrases, chunks, and collocations. The results of corpus linguistics in the works of Swales (1990, 2004), Hyland (2000), Granger, Hung and Petch-Tyson (2002), Steinhoff (2007), Biber and Conrad (2009), Lüdeling and Walter (2009), Römer and Wulff (2010), Nesi and Gardener (2012), and many others have provided us with insights into the linguistic patterns and resources used by certain communities to solve domain-specific rhetorical problems. By using corpus linguistics, language teaching enters a new technological territory with multiple facets that can be applied and tested: (a) strategies of the CALL framework (cf. Beatty, 2003) or (b) Data-Driven Learning (DDL) (cf. Johns, 1986).

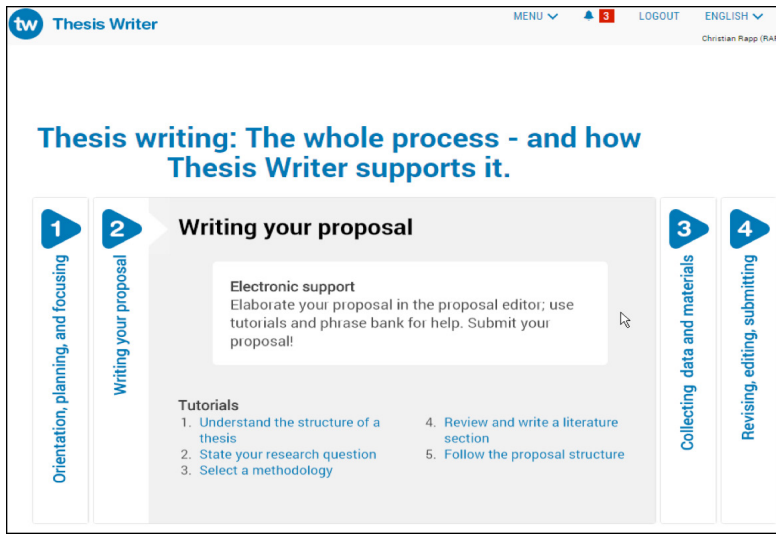
However, technology has scarcely been exploited for interactive tools that support academic writing linguistically (e.g. see Hsieh & Liou, 2009, for a presentation of the POWER and CARE tools). In this study, we will describe several methods of analysis that can be applied to the corpus linguistics results so that they can be used to facilitate academic writing tasks for students writing in English or German (as L1 and/or L2). The methods have been implemented in the interactive academic-writing tool, *Thesis Writer*, designed and constructed jointly by the Department of Applied Linguistics and the Center for Innovative Teaching and Learning of the Zurich University of Applied Sciences in Switzerland. The tool is designed to help students who use either English or German (both as L1 and L2) to write their bachelor or master theses in economics.

2. Method

2.1. Brief description of the academic writing tool

Thesis Writer is primarily a learning platform, but it can also be used as a research tool to collect and analyze data about academic writing. *Thesis Writer* (Figure 1) supports students by (1) structuring the writing process; (2) providing short tutorials for all major steps and actions; (3) offering a "proposal wizard" to guide students through the critical issues of the thesis proposal structure; (4) supporting the transfer of the proposal into the final version of the thesis; and (5) offering help with organizing and revising the thesis.

Figure 1. Road map of *Thesis Writer*



2.2. Technical details

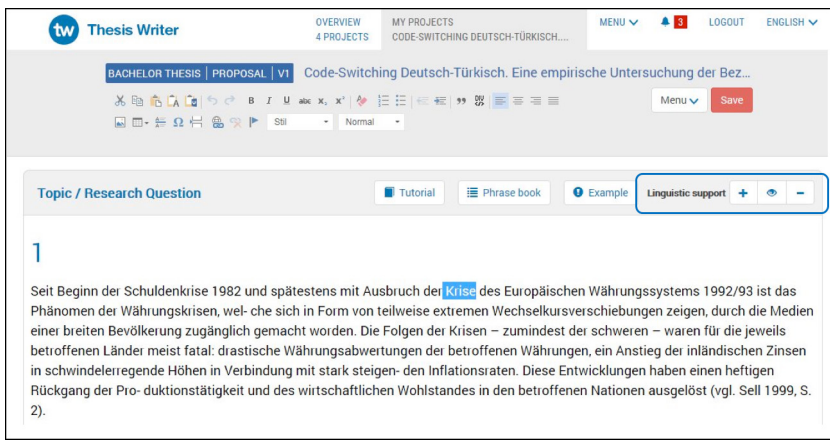
The technical platform for *Thesis Writer* is driven by a LAMP server (Linux, Apache, MySQL, PHP) developed with the PHP-based framework yii1 following strict design patterns for object-oriented programming and the principles of model-view-controller (for more details, see Rapp, Kruse, Erlemann, & Ott, 2015). What happens from a technical perspective when a user seeks language-sensitive linguistic support in *Thesis Writer* by highlighting a word or a passage and clicking the linguistic-support button? The corpus is stored in a database. IMS Open Corpus

Workbench (CWB[4]) enables various queries and actions on the corpus via a number of command line prompts resulting in outputs. To allow the user to perform queries via highlighting and selecting text and using the linguistic support, a Perl script collection and a number of PHP classes mediate between the GUI of *Thesis Writer* and the command line tools of CWB. To improve the quality of suggestions made to the user, we utilize TreeTagger[5] to parse the entire user's text.

2.3. Corpus simple search tool

One of the simplest platform-intermediated corpus methods refers to word-in-context free searches. The IT specialists in the team have helped us design and integrate a user-friendly button, i.e. "linguistic support", so that the linguistic searches are performed directly on the platform by a registered user of *Thesis Writer*, with CWB processing data in the back-end (Figure 2).

Figure 2. Linguistic support tool



2.4. Corpus syntactic search tool

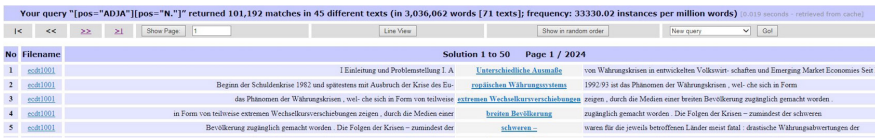
Still in the testing stage, this tool is intended to offer students the option to look for recurrent syntactic patterns, if research demonstrates that such patterns affect the quality of student writing. We looked at [Adj. + Subst.] patterns and found that the syntactic string is quite prolific. One of the challenges at this stage is

4. More information: http://www.ims.uni-stuttgart.de/forschung/projekte/CorpusWorkbench.html

5. More information: Schmid (1994).

also the correct retrieval of the desired POS patterns and the elimination of errors from the list (see the case "schwer -" at the end of the search list in Figure 3). A computational linguist is currently working on solving this matter. The technical solution for the integration in *Thesis Writer* will be implemented by the end of 2015.

Figure 3. Syntactic search in CWB



2.5. Construction of an academic phrase book

A more complex linguistic-support method is the list-of-phrases generator that provides useful academic phraseology when users are working on certain sections of their papers. The phrase book is comparable to the Academic Phrasebank of the University of Manchester[6], but it significantly differs from it because the lists of academic phrases compiled for *Thesis Writer* are organised according to the section of the thesis they are generally typical for. The methodology used for the compilation of the academic phrase book implies several analysis steps:

- *Academic phrases in theory*: Given the fact that the self-compiled corpora are not yet content annotated (e.g. annotation of academic phrases), in order to be able to start the identification of the most frequent academic chunks, a pre-selection stage was performed. This involved the collection of information on academic writing phraseology from textbooks[7] or online informative materials[8].

- *Academic phrases within the research cycle*: Afterwards, we conducted another intermediary processing stage in which the lists of phrases extracted from literature were re-arranged in order to match the sections in *Thesis Writer*: (1) Topic/Research Question, (2) Relevance, (3) Research Gap/ Knowledge Gap, (4) State of the Art, (5) Method/Procedure, (6) Discussion,

6. More information: http://www.phrasebank.manchester.ac.uk/

7. For instance, Bigler and Bugmann (2007).

8. For instance, for academic writing in German, one source of information was bab.la (Schroeter & Uecker, n.d.).

(7) Results, and (8) Conclusions. Each of the sections included sub-categories of phrases as well (see Table 1 below).

- *Keywords in academic phrase lists*: By analyzing the list of phrases resulting from the re-arrangement within the research-cycle categories, we were able to identify several academic keywords.

- *Academic phrase book construction*: Using corpus linguistics methodology, each identified keyword was analysed with the help of a concordance[9] software, which can make instant searches in the self-compiled English and German corpora. Two main strategies led to the identification of the most relevant academic phrases: (a) the software retrieved the clusters in which the indicated "keyword" was included; (b) the analysis was conducted in such a way that the most frequent collocation patterns at the left and right position (+/- 5 words) could be filtered out. From the compiled lists, the most frequent and/or most typical academic-writing phrases were extracted and compiled into a discipline-specific academic phrase book.

Table 1. Academic phrases within the research cycle

Main category in research cycle	**Subcategory in research cycle**	**Keyword(s)**	**Phases (e.g.)**	**Translation EN**
Fragestellung/ Forschungsfrage				*Topic / Research Question*
	Einleitung			*Introduction*
		Arbeit/ Kapitel/Studie/ Abschnit:	- um diese Frage zu beant-worten... - Antwort auf diese Frage - zur Beantwortung dieser Frage - die Frage, ob	*Paper/Chapter/Study/Section:* *- in order to answer this ques-tion…* *- the answer to this question…* *- to answer this question…* *-the question whether…*
		Frage	…	*Question*
		Beginn	…	*Beginning*
	Thema	…	…	*Name topic*
	nennen	…	…	
	…	…	…	...
	…	…	…	…

3. Discussion and conclusion

Although the testing of *Thesis Writer* by users (i.e. students) is still in preparation, there are several hypotheses on which the functionality of the linguistic tools has been based:

9. For simple queries, WordSmith tools (V. 6) (Scott, 2012) were used.

- *Support during writer's block*: It is anticipated that the simple searches will be useful especially to L2 writers during writer's block stages of thesis writing. We imagine that if students have a more or less definite idea of the argumentation line they want to follow at a certain phase of the thesis, they might sometimes have difficulties in identifying the right words/phrases. They then make use of the discipline-specific corpora in order to find out which possible constructions would fit their needs. We do not intend that the students will use this option as a copy-paste procedure, and we would like to prevent that by programming the searches to be retrieved only at a limited left-right number of words.

- *Rhetoric support*: Students sometimes lack the rhetoric awareness of a specific academic genre. TW can help them identify the right argumentative or academic phrase at the time and place they need it.

- *Support for students' writing linguistic diversity*: Scholars often warn against the use of academic phrase lists since it might prevent creativity and encourage repetitions in student writing. However, we anticipate that the diversity of research-cycle-based academic phrases extracted from the corpus (supplemented with the free search in corpus, where students can take inspiration for creating their own repertoire of academic phrases) will be evaluated positively by users.

References

Beatty, K. (2003). *Teaching and researching computer assisted language learning*. New York: Longman.

Biber, D., & Conrad, S. (2009). *Register, genre, and style.* Cambridge: Cambridge University Press. doi:10.1017/CBO9780511814358

Bigler, C., & Bugmann, H. (2007). *Wissenschaftliches Arbeiten: ein Leitfaden für die Ausarbeitung von Masterarbeiten*. Internal report. ETH Zurich. Retrieved from https://www1.ethz.ch/fe/education/teaching_material_secured/Wissenschaftliches_Arbeiten.pdf (login required).

Feilke, H. (2010). "Aller guten Dinge sind drei" – Überlegungen zu Textroutinen & literalen Prozeduren. In I. Bons, T. Gloning, & D. Kaltwasser (Eds.), *Fest-Platte für Gerd Fritz. Giessen.* Retrieved from http://www.festschrift-gerd-fritz.de/files/feilke_2010_literale-prozeduren-und-textroutinen.pdf

Feilke, H. (2012). Was sind Textroutinen? Zur Theorie und Methodik des Forschungsfeldes. In H. Feilke & K. Lehnen (Eds.), *Schreib- und Textroutinen. Theorie, Erwerb und didaktisch-mediale Modellierung* (pp. 1-31). [Forum Angewandte Linguistik Bd. 52]. Frankfurt a.M.: Peter Lang.

Feilke, H. (2014). Argumente für eine Didaktik der Textprozeduren. In T. Bachmann (Ed.), *Werkzeuge des Schreibens. Beiträge zu einer Didaktik der Textprozeduren* (pp. 11-34). Stuttgart: Fillibach bei Klett.

Granger, S., Hung, J., & Petch-Tyson, S. (Eds.). (2002). *Computer learner corpora, second language acquisition, and foreign language teaching*. Amsterdam: Benjamins. doi:10.1075/lllt.6

Hsieh, W.-M., & Liou, H. C. (2009). A case study of corpus-informed online academic writing for EFL graduate students. *CALICO Journal, 26*(1), 28-47.

Hyland, K. (2000). *Disciplinary discourses: social interactions in academic writing*. Harlow, England: Pearson Education.

Johns, T. (1986). Micro-concord: a language learner's research tool. *System, 4*(2), 151-162. doi:10.1016/0346-251X(86)90004-7

Lüdeling, A., & Walter, M. (2009). Korpuslinguistik für Deutsch als Fremdsprache. Sprachvermittlung und Spracherwerbsforschung. Stark erweiterte Fassung von Lüdeling / Walter Korpuslinguistik. In HSK 19, *Deutsch als Fremdsprache*. Mouton de Gruyter, Berlin. Retrieved from http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter innen/anke/pdf/LuedelingWalterDaF.pdf

Myhill, D., & Fisher, R. (2010). Editorial: writing development: cognitive, sociocultural, linguistic perspectives. *Journal of Research in Reading*, *33*(1), 1-3. doi:10.1111/j.1467-9817.2009.01428.x

Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: student writing in higher education*. Cambridge: Cambridge University Press.

Rapp, C., Kruse, O., Erlemann, J., & Ott, J. (2015). Thesis Writer–A system for supporting academic writing. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing (CSCW'15 Companion)* (pp. 57-60). ACM, New York, NY, USA. doi:10.1145/2685553.2702687

Römer, U., & Wulff, S. (2010). Applying corpus methods to written academic texts: Explorations of MICUSP. *Journal of Writing research, 2*(2), 99-127. doi:10.17239/jowr-2010.02.02.2

Schroeter, A., & Uecker, P. (n.d.). bab.la Phrasen–Wissenschaftlich [Online software]. Retrieved from http://de.bab.la/phrasen/wissenschaftliches-schreiben/

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK*.

Scott, M. (2012). WordSmith tools (Version 6) [Stroud: Lexical Analysis Software].

Steinhoff, T. (2007). *Wissenschaftliche Textkompetenz*. Tübingen: Niemeyer. doi:10.1515/9783110973389

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. M. (2004). *Research genres: explorations and applications*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139524827

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; info@research-publishing.net

Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy
Edited by Francesca Helm, Linda Bradley, Marta Guarda, and Sylvie Thouësny

Typeset by Research-publishing.net
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-28-5 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

ISBN13: 978-1-908416-29-2 (Ebook, PDF, colour)
ISBN13: 978-1-908416-30-8 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2015.