International Conference
ICT for Language Learning

Corpus-Based Phraseology Use Within an Academic Writing Platform: a Plagiarism Check

Madalina Chitez[1]

Abstract

Academic Writing and Corpus Linguistics are considered to be major research domains with a high potential for research in applied linguistics and teacher training in higher education. A two-year research project, conducted at the Zurich University of Applied Sciences in Switzerland, has resulted in the creation of an interactive Academic Writing tool, Thesis Writer, which exploits the synergy between these two areas. In order to offer students instant formulation support during thesis writing, two academic-writing corpora have been compiled: TESEC-DE (German) and TESEC-EN (English). The linguistic support tools for thesis proposal writing are: corpus free search and a phrase bank. Academic phrases are distributed according to the section of the thesis students are editing at the time of the online writing process, e.g. Method. In this paper, we will investigate on the actual use of academic phrases and academic vocabulary in students' thesis proposals written in German in the field of Economics. The corpus analysis will include wordlists and concordances extracted from two corpus databases: the reference corpus TESEC-DE and the learner corpus PropCor-DE (Proposal Corpus in German). We will also use a plagiarism detection software program (WCopyfind) to check the degree of phrase "borrowing" from the in-tool corpus in students proposals. Results (appropriate use of academic phraseology, tendency towards routinisation of the proposal outline writing, improved proposal writing process) indicate that academic writing tools are useful instruments for both university students and teachers.

1 Introduction

Academic Writing plays in important role in the achievement of academic success. That is why, the main objective of the two-year research project, "Thesis Writer: A Web-based Learning Environment to Support Dissertation Projects (BA and MA Theses)", conducted jointly by the Department of Applied Linguistics and the Center for Innovative Teaching and Learning of the Zurich University of Applied Sciences in Switzerland, was the creation of Thesis Writer [1]. It offers students the technology-supported learning environment [2] necessary for the writing of their Bachelor and Master Thesis.

On the other hand, one of the major academic writing challenges is plagiarism. It is widely acknowledged that detecting and deterring plagiarism supports both students in the process of fair development of academic writing skills and university tutors in monitoring this process. Nowadays, plagiarism detection is very much facilitated by automated technology. As research reports [3], most plagiarism tools are concerned with external plagiarism detection, which means that student texts are compared with a very large, sometimes dynamic, external database. In fact, most plagiarism software programs intrinsically make use of corpus methodology, since they are designed to identify near-duplicate or identical linguistic chunks (or paragraphs) in a text compared with a reference corpus. In the case of Internet-based plagiarism, the methodology involving natural language recognition techniques is more complex but its aim remains the same: compare a given text with the world-wide-web corpus. It is only few plagiarism software programs which compare two new texts or, at a larger scale, two new text databases, i.e. corpora.

2 Methodology

In the present study, we will combine traditional corpus-analysis methodology [3], mainly using concordance lines, and text-versus-text plagiarism software result analysis. The aim of the study is to identify the degree to which students make use of the *Thesis Writer* linguistic support in their thesis writing process. The procedure is intended to capture two main phenomena: first, how student writers integrate electronic support tool content and indications into their actual texts and, second, whether the integration in texts is natural or it rather raises issues of plagiarism.

[1] Zurich University of Applied Sciences, Switzerland

3 Context and data

3.1 Thesis Writer: an interactive academic writing platform

Thesis Writer is the first technology-supported learning environment that provides an interactive educational platform for writing theses (in German and English). It consists of four major components:

a. Instructions
b. Text-production
c. Training
d. Peer feedback and tutoring

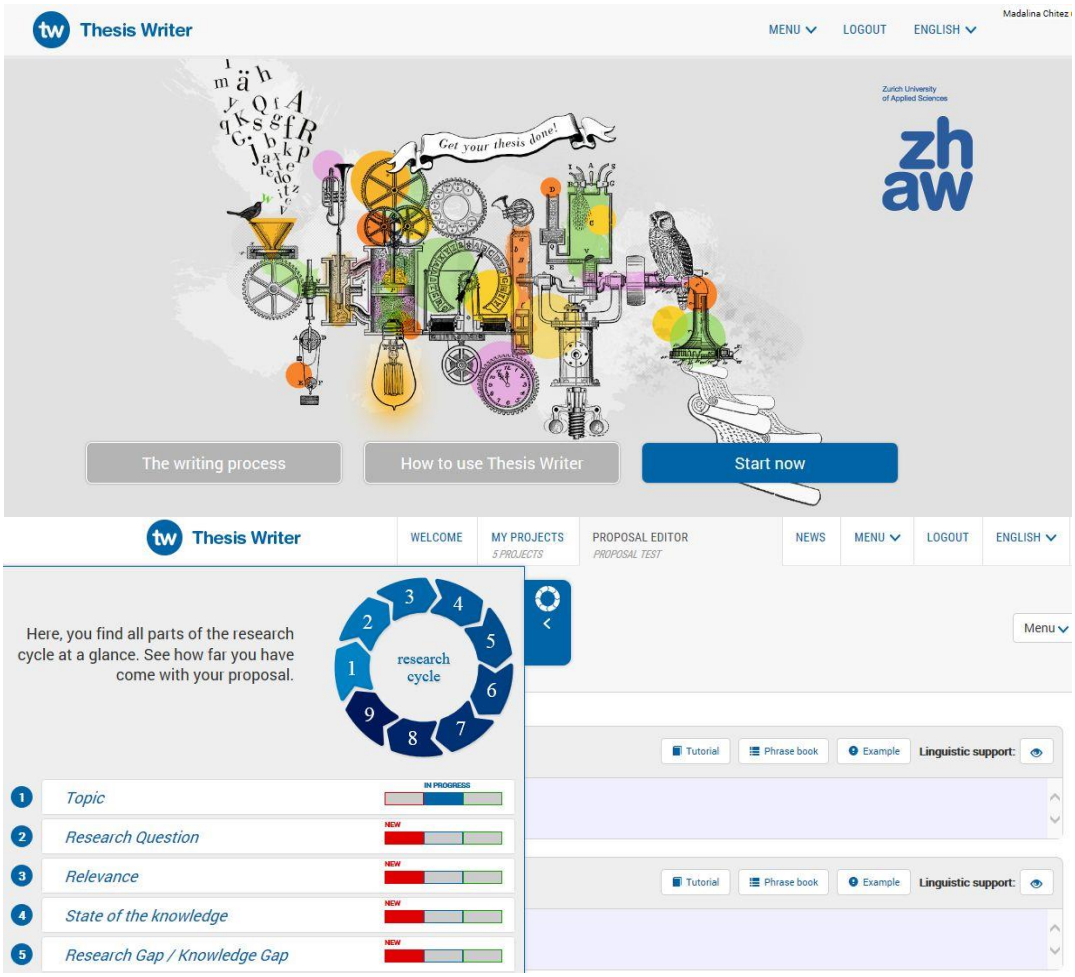

Figure 1: Thesis Writer “Welcome page” and “Proposal-editor”

3.2 The reference corpus: TESEC-DE

The current version of Thesis Writer has been developed and has been successfully implemented in the study program of Economics. For this reason, the linguistic support tools have been designed with the help of a self-complied discipline-specific corpus: TESEC (Corpus of Student Theses in Economics). The size of TESEC corpus reaches 4.781.461 words (2.767.835 w. in TESEC-DE; 2.013.626 w. in TESEC-EN).

3.3 Academic phrases in Thesis Writer

Nine sets of academic phrases can be accessed by students via Thesis Writer. The phrase lists correspond to nine proposal sections: Topic, Research Question, Relevance, State of the Art, Research Gap, Method (Procedure), Results, Discussion, Conclusions. The phrases have been compiled based on information extracted from literature in the field and teaching experience of the tool developers.

3.4 The corpus of Bachelor Thesis proposals: PropCor-DE

The investigation in the present study has been carried out using Bachelor thesis proposals written by the students in the field of Economics at the Zurich University of Applied Sciences with help of Thesis Writer. A total amount of sixteen proposals was collected from Bachelor students during their first semester of Bachelor study. The time period of text writing and collection was March-April 2016. The proposals collected were the result of team-work (4-student teams) and were graded by tutors in the discipline.

The collection of texts has been performed in the context of a didactic exercise of proposal-writing, where the students have been given two main tasks: (a) documentation of the literature review with indication of the resources and inclusion of a scientific-article summary; (b) editing of the proposal based on the general outline in Thesis Writer. Thus, the lecturer, without being an expert in the domain of academic writing, is able to guide the students through the theoretical steps of proposal editing in the discipline he/she is teaching.

4 Analysis

4.1 Data processing

The texts collected in the frame of proposal writing classes, have been processed and transformed into a small analysis corpus, PropCor-DE (Proposal Corpus in German), of 172.004 words. Several steps were necessary in order to clear the data from undesired information which could result in error-prone concordance lists:

1. The transformation of the texts in .pdf format into corpus-specific machine-readable format (.txt);
2. Anonymization and elimination of analysis-disturbing text: e.g. title, personal data, tables, graphs.

Simultaneously with database cleaning, each of the sixteen proposal texts has been assigned a code (e.g. <PropCor_DE_003.1>

4.2 Tools

For the corpus-based analyses, the corpus concordance tool package WordSmith [4] has been used. It allows for a user-friendly and rapid diagnosis of the linguistic phenomena in the database.

As for the additional analyses, one text-versus-text software program has been tested and evaluated for the analysis: WCopyfind (version 4.1.5 freely available at: http://plagiarism.bloomfieldmedia.com/wordpress/software/wcopyfind/).

4.3 Corpus-based phrase-reference control

Each of the phrases included in the Thesis Writer academic phrase lists has been checked against the corpus PropCor-DE.

Section	**Code**	Phrase code / occurrences in PropCor-DE							
Topic	**A**	A1	A2	A3	A4	A5	A6	A7	A8
		3	0	1	4	0	2	2	1

Table 1: *Example of academic phrase check in PropCor-DE*

When concordancing the academic phrases enlisted in Thesis Writer into the corpus database, two strategies were used: either searching for the complete phrase (e.g. *Diese Arbeit beschäftigt sich mit* / EN: This study addresses the issue of) or conducting phrase-triggering word search (e.g. *Ergebnisse* in *Die Ergebnisse der Arbeit bestehen in… /* EN: "Results" in "The results of this paper consist of..."). A manual selection procedure has also been applied, namely the inclusion in the actual-use table only of those academic chunks which have been used in the same proposal section as recommended in Thesis Writer.

4.4 Automatic plagiarism check

A secondary procedure aims at checking whether students have searched for their own phrases using the corpus-based free-search option in Thesis Writer. The analysis inevitably extends towards plagiarism detection in student proposals, considering that they had free access to a discipline-specific expert corpus. For this we needed to compare the reference corpus, TESEC-DE, with the proposal corpus PropCor-DE.

Working with WCopyfind plagiarism detector for research purposes requires additional processing of data: the texts in the two corpora (TESEC-DE and PropCor-DE) had to be aggregated into single .txt documents. The analysis was performed in steps: checking the amount of common string of words for 2-, 3-, 4-, and 5-grams.

4.5 Results

Results of the corpus analysis (example in Table 1) indicate that students have been more tempted to make use of the given phraseology in sections like Topic and Research Gap rather than in sections such as Relevance and Discussion. The results are based on the ratio between used phrases in the proposal sections and the total number of words in PropCor-DE.

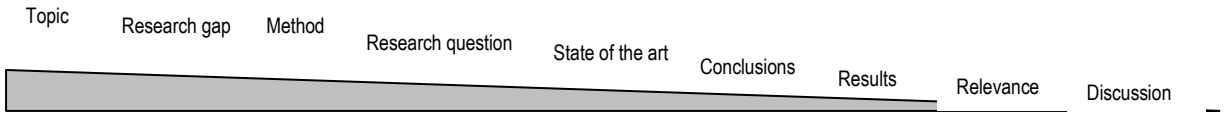

Figure 2: *Intensity of academic-phrase reference in proposal sections*

It is also interesting to note that, the more context-free the phrases are the more students make use of them. When the content or context of the phrase become too specific, e.g. *Unter den Faktoren, die die ...kosten erhöhen, ist … einer der wichtigsten* (EN: Among the factors that increase the costs ... one of the main ones is…), students tend to avoid "borrowing" the expressions in Thesis Writer.
The results of the plagiarism check indicate that most 2-grams (11.487 common strings/cs), 3-grams (5.682 cs) and 4-grams (1.629 cs) are either hazardous word combinations, academic-writing discipline-neutral grouping of words or language-specific collocations: e.g. *aber auch* (EN: but also), *aktuellen Studien* (EN: current studies), *Anforderungen erfüllen* (EN: meet requirements). For 5-grams (391 cs), suspicion can be aroused by the use of certain expressions such as *Ausgaben für Forschung und Entwicklung* (EN: Expenses for research and development) but their concordances in both corpora do not indicate any plagiarism but rather the use of topic-specific vocabulary. Unfortunately, a "borrowing" of the common strings from TESEC cannot be demonstrated. In fact, there are no common strings larger than 5 words.

5 Conclusions

Even if they are study-programme beginners, most students have adopted an expert-like and appropriate outline of the proposal, which has also been positively evaluated by tutors. This proves that Thesis Writer is a very useful tool for preparing Bachelor thesis proposals in point of structure and proposal-section rhetorical appropriacy. At the same time, the phrase-reference from Thesis Writer is not insignificant, with some proposal sections "borrowing" more consistently from the tool phrase lists than others. This means that students do indeed need academic writing routinization support while focusing on content editing. On the other hand, no particular example of extensive "borrowing" from the free-access discipline corpus, i.e. plagiarism, could be detected which means that either students were still not familiar with the corpus free-search option in tool, or they were not convinced of its usefulness, or, in the best case, that they have been correctly initiated in plagiarism rules and followed them accordingly. In order to amend these conclusions, two complementary directions of analysis are necessary: (a) to conduct a screen-log investigation in order to see which tool options led to the incorporation of phrases in writing; and (b) to compile a corpus of thesis proposals in English and compare the use of phrase-reference and plagiarism cases with the data representing the German corpus (displayed in the present case study).

References

[1] Chitez, M., Rapp, C., and Kruse, O. "Corpus-supported academic writing: how can technology help?", Critical CALL - Proceedings of the 2015 EUROCALL Conference, Padova, Italy, F. Helm, L. Bradley, and S. Thouësny. (Eds), Dublin Ireland, Research-publishing.net, 2015, 125-132

[2] O., Erlemann, J., and Ott, J. "Thesis Writer – A System for Supporting Academic Writing", Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing (CSCW2015 Companion), ACM, New York, NY, USA, DOI=10.1145/2685553.2702687 doi.acm.org/10.1145/2685553.2702687, 2015, 57-60

[3] Chitez, M. "Learner corpus profiles: the case of Romanian Learner English", Linguistic Insights Series (Series Editor: Maurizio Gotti), Bern/Berlin/Bruxelles/Frankfurt am Main/New York/Oxford, Peter Lang, 2014.

[4] Scott, M. "Lexical Analysis Software WordSmith Tools version 6", Stroud, 2012.