

Sebastian Jan Dunat

ACADEMIC LANGUAGE FUNCTIONS IN CONFERENCE ABSTRACTS CORPUS ANALYSIS OF ACADEMIC LANGUAGE FUNCTIONS

STRESZCZENIE

3.05.2025r.

Key words: academic language, academic language functions, abstracts, corpus analysis, corpus analysis of abstracts, corpus studies, gender based corpus studies

There are various ways to present, define and analyze the structure and rhetorical functions of abstracts, as the academic studies mentioned in this dissertation illustrate. Although previous studies have addressed many purposes of research on abstracts, not only in terms of corpus linguistics, none have focused on academic language functions in abstracts, especially using computational linguistics methods for research. The present dissertation is organized in the following order: the material is assessed and described with the use of a special corpus and computational linguistic methods to compare the academic vocabulary used in conference abstracts by both sexes in all of the characterized academic fields. The methodology of Averil Coxhead (2000) is used to establish the most frequent vocabulary data in question. The analysis aims at detecting possible interdisciplinary variation in the frequency of occurrence of academic language functions. Gender is established as the leading parameter of the study. The presented corpus data, used for lexical analysis, is labeled, firstly into parts of speech, and secondly into three categories: Vocabulary of the research process, Vocabulary of analysis, and Vocabulary of evaluation, based on Martin (1976). The corpus data is chosen from five books of abstracts from five conferences, each on a different continent. It is selected by hand to equalize the number of examples, according to the gender factor.

The corpus data is divided into three academic fields: Computational Statistics, Humanities, and Biological & Health Sciences. The corpus has been designed in the present study to determine the quantitative and, to some extent, qualitative differences in the use of academic language functions in various fields of written academic language. The data of the corpus provide the following elements: Source, Reference number and page number, Author, Gender, Speaker type, Institution, Speaker origin, Discipline, Title, Abstract text, Abstract tagged for the parts of speech, and Number of nouns, verbs, adjectives, and adverbs used in the corpus texts. It is expected that there will be differences in the realization of language functions and academic vocabulary in the groups under analysis.

The thesis is divided into five chapters. The first presents a theoretical background of

communication in academic discourse, the genre of abstracts, corpus, and principles of computational linguistics, gender aspect in linguistic research, data evaluation in statistics, and academic vocabulary and functions of academic language. This discussion, comprised in five sections of the first chapter, forms the basis of the introduction of important notions used for the implementation of this dissertation. The second chapter consists of descriptions of practical approaches to abstract analysis, not only in corpus studies. The discussion focuses on seven studies reviewing presented journal articles, which underline cross-linguistic, disciplinary, intercultural, and diachronic perspectives as well as rhetorical structure, and corpus study elements. The third chapter presents the analysis design and introduces computational means of the analysis, i.e., software programs used in the dissertation (R, AntConc, ProtAnt, TagAnt). The central part of the thesis, the fourth chapter, provides an analysis of the corpus data, researched in terms of gender and speakers' origin to visualize better the differences and similarities occurring in the researched material. The quantitative computational approach is the primary method of analyzing the chosen data samples, i.e., the interpretation of the presented tables, considering the female and male usages of specified lexis, and the quantitative use of specified academic language functions, is verified by Pearson's Chi-square test conducted in the R program, and the Ant software tools. The last part of the dissertation contains conclusions corresponding to the data interpretation presented in each section of the fourth chapter and suggestions for further investigation. It offers a survey of the similarities and differences of the research data for gender and writers' native/non-native factors. The outcomes of the study are discussed in reference to selected sources. Conclusions are backed up by Pearson's Chi-Squared Test (where appropriate), extracted from the R software, visualizing the numerical differences between the specified data.

The chi-square tests confirmed several statistically significant relationships across different academic fields, with variations based on gender and writers origin. Firstly, the application of the F7 function, related to evaluative processes, exhibited statistically significant differences between the examined disciplines and was found to be influenced by gender ($p=0.02$). Secondly, the chi-square test identified a notable statistical association ($p=0.01$) regarding the use of the F9 function (organizing order of information) in relation to the total number of native and non-native writers using this function in abstracts across the studied fields. Thirdly, the F1 function, related to research processes, was analyzed in reference to the combined parameters of the academic field and the origin of the writers, yielding a p-value of 0.009. Additionally, the F3.1, F3.2 and F3.3 functions, related to

presenting methodology, examined with respect to the combined parameters of gender and academic field, demonstrated statistical significance at p-values of 0.008, 0.03 and 0.004, respectively. Moreover, the F4 function (reporting results) was found to be more frequently used by native writers in the field of Biology and Health, whereas non-native writers employed the F4 function more frequently in the remaining fields analyzed ($p=0.03$). Finally, the F5 function, related to presenting conclusions, yielded a p-value of 0.002, which indicates a statistically higher frequency of the F5 usage by non-native writers in the disciplines of Computer Statistics and Humanities found in the studied material.

Sebastian Jurek