A Study of Scientific Writing: Comparing Theoretical Guidelines with Practical Implementation

Mark Kröll* Know-Center GmbH Graz, Austria mkroell@know-center.at

Gunnar Schulze* Know-Center GmbH Graz, Austria gschulze@know-center.at

Roman Kern Know-Center GmbH Graz, Austria rkern@know-center.at

Abstract

Good scientific writing is a skill researchers seek to acquire. Textbook literature provides guidelines to improve scientific writing, for instance, "use active voice when describing your own work". In this paper we investigate to what extent researchers adhere to textbook principles in their articles. In our analyses we examine a set of selected principles which (i) are general and (ii) verifiable by applying text mining and natural language processing techniques. We develop a framework to automatically analyse a large data set containing $\sim$14.000 scientific articles received from Mendeley and PubMed. We are interested in whether adhering to writing principles is related to scientific quality, scientific domain or gender and whether these relations change over time. Our results show (i) a clear relation between journal quality and scientific imprecision, i.e. journals with low impact factors exhibit higher numbers of imprecision indicators such as number of citation bunches and number of relativating words and (ii) that writing style partly depends on domain characteristics and preferences.

1 Introduction

Writing good scientific articles is a skill. Researchers seek to acquire this skill for the purpose of successfully disseminating their ideas to the scientific community. Learning to write good articles is a process that for most of us starts at graduate level and keeps us company in the course of our careers. To advance the learning process, there is (i) plenty of literature out there containing do's and dont's, (ii) seniors administering doses of advice and (iii) entire lectures dedicated to this very subject.

In this paper, we investigate whether researchers do adhere to general writing principles taken from textbook literature. We are interested in whether adhering to writing principles is related to the journal quality, the scientific domain or gender and whether there is a change over time. Doing so allows us to better understand which and to what extent theoretical guidelines are practically implemented. Deviations from textbook literature could be indicators of good practice and if they occur frequently enough, they might also be candidates for textbook updates.

Studying current trends in academic writing (cf. (Tas, 2010)) originates in the domains of pragmatics and linguistics. In this research area we recognize two larger directions. The first one seeks to relate an article's content to scientific concepts, for instance, whether an article contains a theory or not (cf. (Pettigrew et al., 2001)) or to scientific discourse elements, for instance, which paragraphs can be related to categories such as Motivation or Experiment (cf. (Liakata et al., 2012)). The other direction focuses more on organisation and structure including the analysis of entire scientific theses (cf. (Paltridge, 2002)) or the analysis of single structural elements such as the title (cf. (Soler, 2007), (Haggan, 2003)).

In contrast to previous work, we conduct our analyses at a larger scale. We thus develop a framework to automatically analyze large amounts of scientific articles. In our experiments we select writing principles which are on the one hand general and often recommended in textbook literature (cf. (Lebrun, 2007), (Alley, 1996)) and on the other hand automatically retrievable and verifiable by applying text

* These two authors contributed equally to this work.

Proceedings of SADAATL 2014, pages 48–55,
Dublin, Ireland, August 24, 2014.

mining and natural language processing techniques. To give an example, the principle "use active voice when describing your own work" is a popular one and can be verified by examining the verb types in the article's abstract, introduction and conclusion. In our study we analyze two data sets - one from Mendeley[1], a popular reference management tool, - one from PubMed[2], a free resource containing citations for biomedical literature. We can observe relations between journal quality and textbook recommendations such as *Avoid Imprecision* and *Engage the Reader*. In addition, the results indicate writing style preferences due to domain characteristics. Our findings show that theoretical guidelines partly concur with practical implementation and thus contribute to better understand the extent to which theory guides praxis and vice versa praxis might guide theory.

The remaining paper is organized as follows: Section 2 provides details on the used data sets as well as the software framework to automate the analysis of scientific articles. Section 3 contains experimental results and discussions of analyzed writing principles. Related work is covered in Section 4 and concluding remarks are presented in Section 5.

2 Experimental Setup

2.1 Data Sets

For our analyses we used scientific articles from two sources - Mendeley and PubMed - which also provided us with meta data, e.g. name of the conference or journal. Most of the publication organs were journals and we decided to select only journals which had a minimum of 10 articles and for which we could find a respective 5-year impact factor[3]. We decided to conduct our analyses over a 10-year time period from 2001 to 2010, since only in this period articles from both sources were available. In total we experimented with 13866 scientific articles. Grouping them according to scientific quality, domain and gender, we constructed three data sets described in the following:

- Quality: According to the impact factor (IF), we divided the scientific articles into three groups; low IF ranging from 0 to 2.5 (2303 articles), middle IF ranging from 2.5 to 4 (5734 articles) and high IF ranging from 4 to 35 (5829 articles). The ranges were chosen to reflect the journal quality while containing an appropriate (not too small) number of scientific articles per category.

- Domain: We divided the scientific articles by their journal type into two groups: biomedical (7053 articles) and a technical (6813 articles) which contained mainly articles from physics and computer science.

- Gender: We used two gazetteer lists to identify female[4] or male[5] first authors. Since only a part of the authors' first names was unabbreviated, we used a subset of articles for these experiments: number of articles with male first author = 1182, number of articles with female first author = 1990.

2.2 Framework

To automatically analyze large amounts of scientific articles, we designed a framework and embedded our analysis algorithms in a Hadoop[6] environment. The environment allows parallelization of processes and thus greatly reduces computation time. We stored the results in a PostgreSQL[7] database for quick access and used various Python packages such as matplotlib[8] for creating graphical representations of our results.

Our first pre-processing step encompassed the extraction of textual content from scientific publications. To automatically extract the content, we used a processing pipeline (cf. (Klampfl et al., 2013))

[1] http://www.mendeley.com/
[2] http://www.ncbi.nlm.nih.gov/pubmed
[3] http://www.citefactor.org/impact-factor-list-2012.html
[4] http://deron.meranda.us/data/census-dist-female-first.txt
[5] http://deron.meranda.us/data/census-dist-male-first.txt
[6] http://hadoop.apache.org/
[7] http://www.postgresql.org/
[8] http://matplotlib.org/

that applies various machine learning techniques in combination with heuristics to detect the logical structure of a PDF document. Further processing steps included (i) tokenization, (ii) sentence splitting, (iii) stemming, (iv) part-of-speech tagging and (v) chunking. We employed part-of-speech and chunking information in our analyses (see Section 3) to distinguish verb phrases with respect to present vs. past tense as well as active vs. passive voice.

3 Analysis of Scientific Literature

In this section we analyze a set of selected writing style principles with respect to *Reader Engagement* and *Imprecision*. Each analysis contains (i) a motivating statement mostly taken from (Lebrun, 2007), (ii) a visual representation of results and (iii) an interpretation of results. During our experiments we could observe that most of the time there were no significant differences between articles written by male and female first authors. We repeated the experiments with a majority criterion of authors, i.e. more female first names or more male first names per article, resulting in similar findings. It appears that both genders adhere to the same guidelines which were standardly used at the time.

3.1 Engaging the Reader

In this section we examine different means to engage the reader according to textbook literature including (i) the title, (ii) figures & tables and (iii) a lively writing style based on using present tense and active voice.

3.1.1 Title

The title represents the first point of contact with the reader (and the reviewer) and should ideally be made catchy and standing out. We examine three means to do that: (i) usage of verbs to increase energy, (ii) usage of acronyms to provide a reference shortcut for others and (iii) usage of questions to create a hook. Figure 1 contains average numbers of article titles with respect to these means.

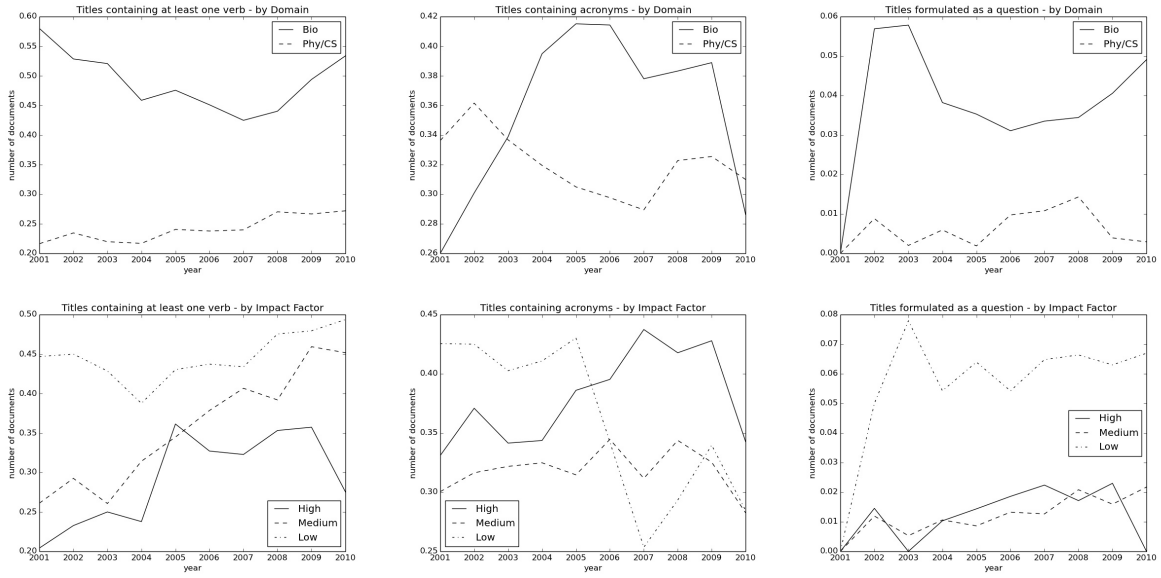

Figure 1: Illustration of (i) titles containting at least one verb (left), (ii) titles containing acronyms (middle) and (iii) titles which contain a question (right) over a ten-year time period. The upper row reflect distinction by domain, the lower row by impact factor. The y-axis represents the average number of article titles exhibiting the respective feature.

The upper left figure in Figure 1 tells us that using verbs in titles is more common among authors in the biomedical domain than in the physics/computer science domain. The lower left figure indicates a trend towards using more verbs in the title over the years independent of the journal quality. The upper, middle figure of Figure 1 shows that using acronyms in titles is more common in the biomedical domain

and a possible trend using acronyms at the beginning of the century. The lower figure in the middle indicates an up and down over the years across impact factors. The right figures in Figure 1 tell us that only a low percentage of authors use questions in their titles independent of domain or journal quality.

The numbers corroborate textbook literatures' recommendation of using verbs in the titles as well as using acronyms. A bit surprising is that questions in titles are rarely used, since according to literature they create a mighty hook for the reader. In a next step we intend to relate the title to the content of the abstract and the introduction to answer the question how well the title reflects the article's content.

3.1.2 Figures & Tables

Visual representations of results in terms of figures and partly of tables help the reader to reduce reading time. According to (Lebrun, 2007) they even represent visual information burgers which are easy to digest. Figure 2 contains respective average figure and table counts.

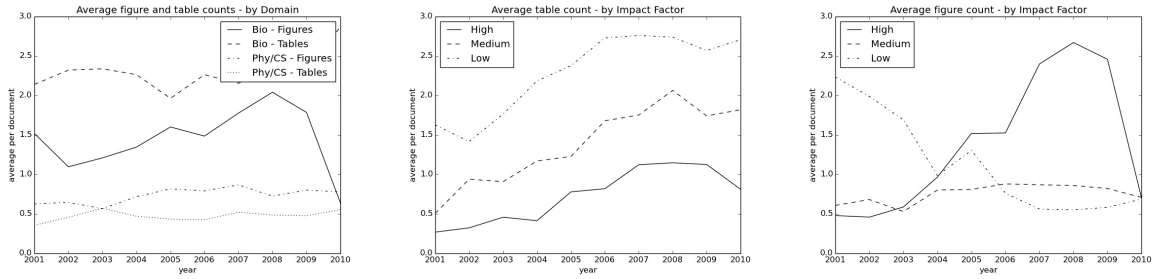

Figure 2: Illustration of average figure and table counts over a ten-year period according to domain (left) and impact factor (middle, right). The y-axis represents the average number of tables/figures per article.

The left figure in Figure 2 illustrates that authors from the biomedical domain use more tables and figures than authors from the physics/computer science domain. The middle and right figure reflect average counts according to impact factor. Journal articles with a high impact factor contain (i) fewer tables than journals with middle or low impact factors and (ii) in general more figures.

From the results in Figure 2 we learn that usage of figures and tables appears to a certain degree be dependent on the domain. In biomedicine, the usage of figures to convey information seems more widespread than in technical domains. We assume even higher figure counts in domains such as chemistry where illustrations, for instance, of molecules are far more frequent. In addition, it seems that authors of high impact journals prefer using figures to using tables probably because the information content is more easily to grasp. Tables appear to be more suited to structure information. In a next step we also intend to analyze figures' and tables' captions with respect to comprehensiveness, i.e. to what extent are captions self-contained?

3.1.3 Lively Writing Style

Textbook literature advices authors to formulate their contributions in an active way using active voice and the present tense. To learn more about present tense usage, we simply counted the occurrences of the respective part-of-speech tags[9], i.e. VB, VBP and VBZ. To count occurrences of active voice, we inspected all identified verb chunks whether they contained auxiliary verbs as well as a past participle part-of-speech tag. If they did, we considered them passive voice otherwise active voice. Figure 3 contains average fractions of verb phrases with respect to present tense and active voice.

The upper left figure in Figure 3 illustrates that authors of the physics/computer science domain use a lot more present tense compared to authors from the biomedical domain. The upper right figure indicates that the higher the journal's impact factor the more present tense is used by the authors. The lower left figure indicates that active and passive voice are almost evently distributed throughout article contents with a bit passive predominance. The lower right figure shows no significant difference of using active voice with respect to journal quality. There is but a trend towards using more active voice over the years.

[9] http://www.cis.upenn.edu/~treebank/home.html

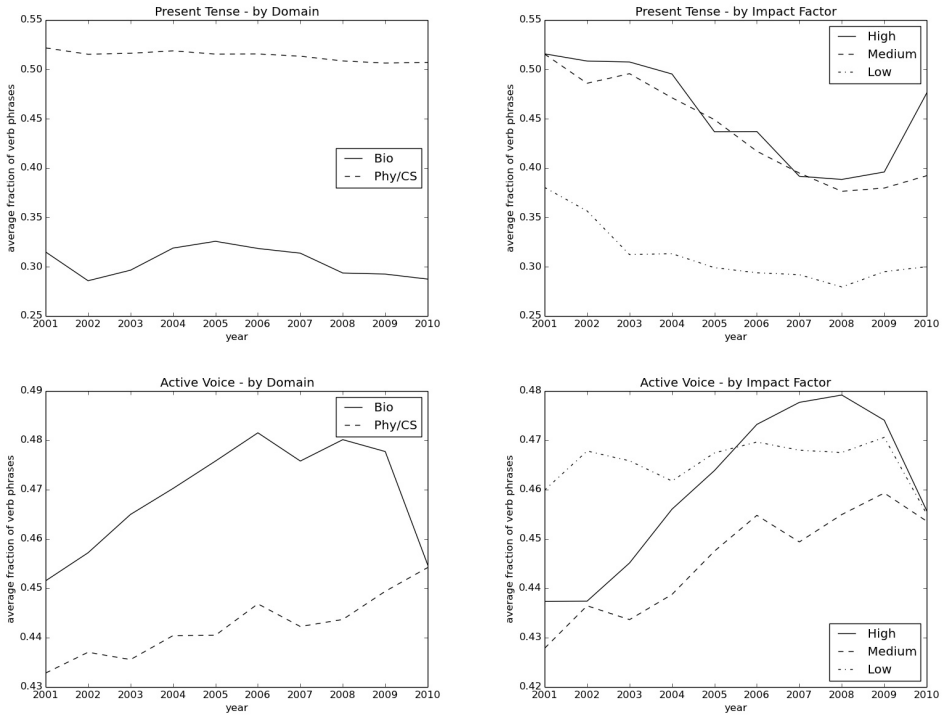

Figure 3: Illustration of average fractions of verb phrases with respect to present tense (upper figures) and active voice (lower figures) over a ten-year period. The left figures correspond to analyses with respect to domain and the right ones to analyses with respect to impact factor.

The observed high percentage of present tense and active voice verb phrases adheres to the textbook principle of lively stating one's own work. Yet, in this analysis we took into account the tense and the voice for the entire article content. To address this issue in greater depth, we intend to solely analyze abstract, introduction and conclusion in the near future.

3.2 Imprecision

In this section we examine indicators of scientific imprecision according to textbook literature including (i) number of citation bunches and (ii) number of relativating words.

3.2.1 Citation Bunches

Statements such as "many people have been working in this research area" + a sequence of citations indicate lack of precision or insufficient dealing with the subject matter. We define a collection containing more than 3 citations as *citation bunch*. Figure 4 contains average numbers of citation bunches per article.

The left figure in Figure 4 indicates that citation bunches occur more often in the biomedical domain than in the physics/computer science domain. The right figure shows that citation bunches occur far more often in journals with a low impact factor than in those with a middle or high one.

The findings indicate that journals with a higher impact factor contain fewer citation bunches - one indicator of lack of precision. Concerning the higher numbers in the biomedical domain we intend to examine the type of scientific articles in the future; for instance, we assume that survey articles contain more citation bunches than others.

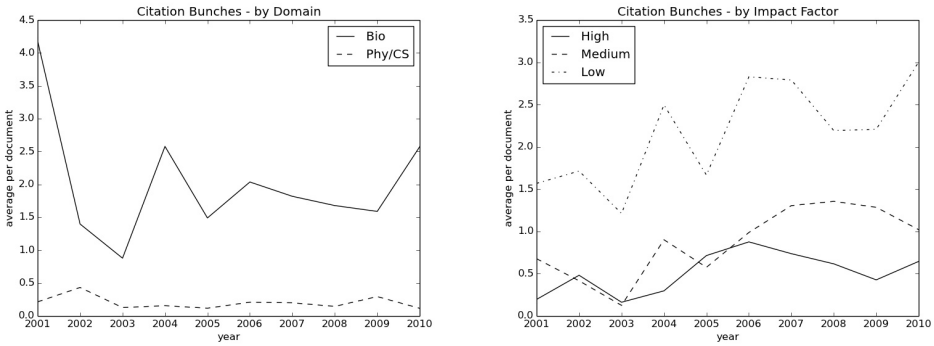

Figure 4: Illustration of averaged citation bunch counts according to domain (left) and impact factor (right)) over a ten-year period. The y-axis corresponds to the average number of citation bunches (>3 citations) per scientific article.

3.2.2 Usage of Relativating Words

Overusage of relativating words[10] indicates lack of precision. Reviewers may doubt an author's expertise and assurance of results if relativating words occur too frequently. Figure 5 contains average numbers of relativating words per article sentence.

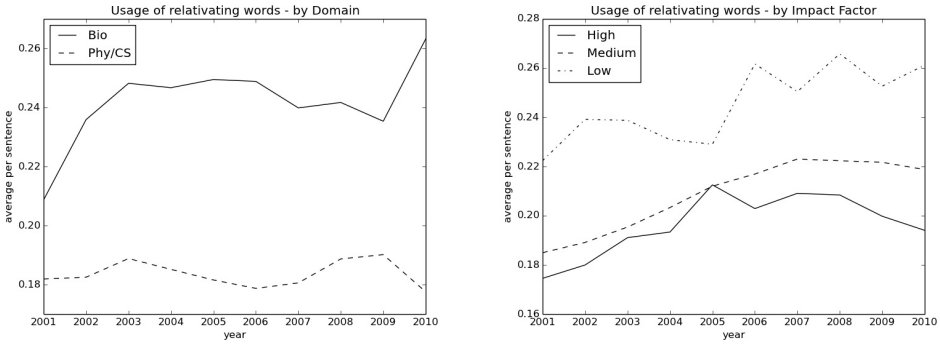

Figure 5: Illustration of relativating words usage by domain (left) and by impact factor (right). The y-axis represents the average number of relativating words per article sentence.

The left figure in Figure 5 shows that more relativating words are used in the biomedical domain than in the physics/computer science domain. In the right figure journals with a high and a middle-ranged impact factor exhibit fewer relativating words than journals with a low impact factor.

The higher usage of relativating words in the biomedical domain remains unclear and might be due to domain characteristics. Regarding the journal quality a similar behavior as in Section 3.2.1 can be observed: higher quality journals contain fewer relativating words - an indication for a higher scientific preciseness and quality.

4 Related Work

The analysis of academic writing originates from research areas such as linguistics and pragmatics. These areas are rather interested in studying *how* scientific articles are written instead of *what* kind of knowledge they contain.

Discourse Analysis is a modern discipline that studies amongst other things language beyond the level of a sentence taking into account the surrounding contexts as well. The detection of discourse structure in scientific documents is important for a number of tasks such as information extraction or text

[10] According to (Lebrun, 2007) relativating words include significantly, typically, generally, commonly, may/can, a number of, the majority of, substantial, probably, several, less, various, frequent, many, others, more, often, most, a few, the main.

summarization. Elements of discourse include the statement of facts, claims and hypotheses as well as the identification of methods and protocols. In this context (Liakata et al., 2012) automate the recognition of 11 categories including Hypothesis, Motivation and Result to access the scientific discourse of scientific articles. In the Partridge system, (Ravenscroft et al., 2013) build upon the automated recognition to automatically categorize articles according to their types such as Review or Case Study. (Teufel et al., 2002) used discourse analysis to summarize scientific papers. She restored the discourse context by adding the rhetorical status, for example, the scientific goal or criticism, to each sentence in an article. In a similar way, (Liakata et al., 2013) take scientific discourse into account to generate a content model for summarization purposes.

Besides analyzing the structure and organisation of entire publications (cf. (Paltridge, 2002)), there is related literature dedicated to the analysis of single (structural) elements including (i) the title or (ii) citations. The title is of particular importance often representing the first point of contact with the reader. (Haggan, 2003) investigated whether titles of scientific articles could be regarded as headlines with a clear role of informing and engaging the reader. In her work she pointed out the relation between title formulation and information advertisement. (Soler, 2007) conducted title studies in two genres (review and research papers) and in two fields (biological and social sciences). She statistically analyzed titles with respect to word count, word frequency and title construction.

Citation analysis represents one of the most widely used methods of bibliometrics which aims to quantitatively analyze academic literature. Citation analysis (cf. (Garfield, 1979)) is an expression for simply counting a scientific article's citations which can be regarded as indicator for an article's scientific impact; the more often the article is cited, the higher its academic value (cf. (Garfield, 1972)). An important part of citation analysis represents hedge detection (cf. (Lakoff, 1972)). Hedges are linguistic devices which indicate that authors do not or cannot back up their statements with facts. Hedge detection, thus, supports the distinction between facts and unreliable or uncertain information (cf. (Crompton, 1997)). Facing the continuously growing amounts of scientific articles there has been an increased interest in automating the process (cf. (Di Marco, 2006), (Farkas et al., 2007)).

5 Conclusion

Our paper's contribution encompasses a comparison of theoretical guidelines, i.e. "What the literature recommends?" with their practical implementations, i.e. "How authors actually write scientific articles?". We designed a framework to automatically analyze $\sim$14.000 scientific articles with respect to a selected set of writing principles.

To summarize the results: Section 3.2 shows a clear relation between journal quality and imprecision, i.e. journals with low impact factors exhibit higher numbers of imprecision indicators such as number of citation bunches and number of relativating words. In addition, the number of figures and the percentage of verb phrases in present tense tend to be higher with higher quality journals (see Section 3.1).

In respect to the domain, the results indicate writing style preferences probably due to domain characteristics, for instance, usage of more figures (see Section 3.1.2) and domain preferences, for instance, lesser usage of present tense (see Section 3.1.3).

Other interesting observations include (i) that adhering to writing principles appears to be gender independent and (ii) that using acronyms in titles is far more popular than using questions in the title (see Section 3.1.1) independent of domain and impact factor.

Our findings show that theoretical guidelines partly concur with practical implementations and thus contribute to better understand the extent to which theory guides praxis. A better understanding will contribute (i) to confirm textbook principles and (ii) to update writing principles due to good practice. In a next step we plan to extend the scale of our analyses to include several hundred thousand scientific articles as well as the complexity of our analyses to investigate issues including (i) paper skeleton, for instance, "Is there a prefered heading structure?" and (ii) usage of synonyms which hampers clarity.

Acknowledgements

We thank Mendeley for providing the data set as well as Werner Klieber for crawling the PubMed data set. The presented work was developed within the CODE project funded by the EU FP7 (grant no. 296150). The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

[Alley1996] Alley, M. 1996. *The Craft of Scientific Writing*. Springer.

[Crompton1997] Crompton, P. 1997. *Hedging in academic writing: Some theoretical problems.* English for Specific Purposes 16 (4).

[Di Marco2006] Di Marco, C., Kroon, F. and Mercer R. 2006. *Using Hedges to Classify Citations in Scientific Articles.* Computing Attitude and Affect in Text: Theory and Applications. Springer Netherlands.

[Farkas et al.2007] Farkas, R., Vincze, V., Mora, G., Csirik, J. and Szarvas, G. 2010. *The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text.* Proceedings of the Fourteenth Conference on Computational Natural Language Learning.

[Garfield1972] Garfield, E. 1972. *Citation analysis as a tool in journal evaluation.* Science (178).

[Garfield1979] Garfield, E. 1979. *Citation Indexing: Its Theory and Applications in Science, Technology, and Humanities.* John Wiley, New York, NY.

[Haggan2003] Haggan, M. 2003. *Research paper titles in literature, linguistics and science: dimensions of attraction.* Pragmatics 36 (2).

[Lakoff1972] Lakoff, G. 1972. *Hedges: A study of meaning criteria and the logic of fuzzy concepts.* Papers from the Eighth Regional Meeting, Chicago Linguistics Society Papers.

[Lebrun2007] Lebrun, J. 2007. *Scientific Writing*. World Scientific Publishing Co Pte Ltd.

[Liakata et al.2012] Liakata, M., Saha, S., Dobnik, S., Batchelor, C. and Rebholz-Schuhmann, D. 2012. *Automatic recognition of conceptualization zones in scientific articles and two life science applications.* Bioinformatics 28 (7).

[Liakata et al.2013] Liakata, M., Dobnik, S., Saha, S., Batchelor, C. and Rebholz-Schuhmann, D. 2013. *A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task.* Proceedings of the Conference on Empirical Methods in Natural Language Processing.

[Klampfl et al.2013] Klampfl, S. and Kern, R. 2013. *An Unsupervised Machine Learning Approach to Body Text and Table of Contents Extraction from Digital Scientific Articles.* Research and Advanced Technology for Digital Libraries.

[Paltridge2002] Paltridge, B. 2002. *Thesis and dissertation writing: an examination of published advice and actual practice.* English for Specific Purposes 21 (2).

[Pettigrew et al.2001] Pettigrew, K. and McKechnie, L. 2001. *The use of theory in information science research.* American Society for Information Science and Technology, 52.

[Ravenscroft et al.2013] Ravenscroft, J., Liakata, M. and Clare, A. 2013. *Partridge: An Effective System for the Automatic Classification of the Types of Academic Papers.* AI-2013: The Thirty-third SGAI International Conference.

[Rubin2004] Rubin, R. 2004. *Foundations of Library and Information Science.* 2nd ed. New York: Neal-Schuman.

[Soler2007] Soler, V. 2007. *Writing titles in science: An exploratory study.* English for Specific Purposes 26 (1).

[Tas2010] Tas, E. 2010. *"In this paper I will discuss": Current trends in academic writing*. Procedia - Social and Behavioral Sciences.

[Teufel et al.2002] Teufel, S. and Marc Moens, M. 2002. *Summarizing scientific articles: experiments with relevance and rhetorical status*. Computational Linguistics 28 (4).