Notes on Similarities and Differences between Abstracts from a Local and a Top-Ranked Computer Science Journal from a Text-Analysis Point of View

Assoc. Prof. Rodica Ioana Lung

Faculty of Economics and Business Administration, Babes-Bolyai University of Cluj Napoca, Romania
rodica.lung@econ.ubbcluj.ro

Doi:10.5901/mjss.2015.v6n1s1p82

Abstract

Aim This paper presents a comparison between abstracts of papers published in a national level journal from Romania and one considered top - ranked in computer science with the aim of identifying possible common features or differences. Method The comparison is based on a registry variation text analysis that computes potential indicators of scientific writing (number of nouns, verbs, and adverbs used, lexical density and type - token ratio) Results Results show a significant difference in terms of number of nouns, adverbs and verbs used, indicating that there is a need for the Romanian researchers to take into account some aspects related to scientific writing (particularly the use of nouns in academic texts). Conclusion While this analysis does not aim to account for any qualitative differences regarding content between the two journals, in an era in which open-access publication may change the way article impact is measured, it does emphasize the importance of academic writing skills development and of targeted academic writing programs.

Keywords: *academic writing, writing in the disciplines, text analysis*

1. Introduction

Academic writing has emerged as a research field from the intrinsic need of researchers to communicate their results in the most efficient manner possible, both in terms of clarity of exposure and impact (measured in terms of the size of the audience reached). In the very competitive world of scientific publishing academic writing skills are a powerful tool that can make the difference between “accept” and “reject”, regardless of the scientific content of the paper. Journal rankings extend their validating power not only to article contents but also to the manner articles are written; academic writing courses for graduate students usually promise support for writing articles for high-impact journal papers.

In such a competitive setting, Romanian universities and affiliated journals struggle to provide their researchers the tools for disseminating their research. The current open-access trends and the plethora of indexing databases offer the illusion that online publishing ensures visibility and impact for the research. Out of the article indexing systems, the Thomson ISI Journal Citations Reports is the one used in Romania for all research related evaluations: grant proposals, promotions, tenure, etc. In this context, local (national) journals are becoming a forum for dissemination where PhD students and young researchers are training themselves to the art of academic writing; in most cases, this is the only kind of self-training they can get, as formal support for scientific writing skills development is still scarce.

In this context, the question raised by this paper is: how different are the papers published in a local journal from those from a top-ranked one in the same field? Are there any quantitative indicators that can measure the degree of scientificity of a text and if so, how do the texts differ? Corpora analysis methods (Conrad 1996, Mihalcea, Corley & Strapparava 2006) do offer some solutions to this problem: scientific writing presents some specific features that distinguish it from other types of texts: a higher number of nouns (expressing concepts, terms, etc.) and lower number of verbs and adverbs (Bieber 1993, 2006). Other potential indicators are the type-token ratio and the lexical density. We have used these indicators to analyze abstracts from two journals, a local (national) one affiliated to a university, and a top-ranked one; results show that there are significant differences as far as the indicators analyzed are concerned.

2. Background

Various abstract analyses have been performed and reported for different purposes. Most existing studies emphasize both the importance of the abstract and the fact that they represent a special type of writing that has to be acknowledged. Cross & Oppenheim (2006) present an analysis of abstracts structure in an attempt to identify modifications in abstract styles written by the authors themselves instead of editors of a protozoology journal. In their paper they point out a number of reasons abstracts are important and should follow the specified guidelines.

The reasons mentioned by them include: abstracts save readers time – the decision if the paper contains the information they look for is made based on the it; in some cases it can provide some language preparation through keywords and ideas; and a well written abstract can present the full argument of the original article. Obviously, authors should keep in mind these reasons as motivation for constructing their abstracts in a reader-oriented manner.

In spite of author's knowledge about the importance of abstracts, it does happen that abstracts do not meet these requirements. Tibbo (1993) performed a study of abstracts from various disciplines pointing out repercussions on the efficiency of abstracting and indexing systems. This further affects the research process as scientists rely on scientific databases to document their research and select papers to fully read based only on the abstract listed there.

It is accepted thus that both writing and reading abstracts is not trivial: they should be short and contain the main arguments of the article. It follows that the contents should be *lexically and propositionally dense* (Hartley, 1994; Kaplan et al., 1994).

Another extensive study examined 90 abstracts in three applied linguistics journals (Tseng, 2011) from two points of view: the move structure and verb tense in each move. They report that the studied abstracts tend to follow only a four move structure instead of the recommended five – justifying it with the size limit. However, the five move pattern recommended in applied linguistics (Background, Aim, Method, Results, and Conclusions – Swales & Feak 2004, Santos 2009) is specifically designed for "short" abstracts. The four moves used were Aim, Method, Results, and Conclusions leading to the conclusion that the Background may have been considered optional. However, the Background it is probably the move most difficult to express in the short and concise manner required in writing an abstract. Regarding the verb tense the authors report the present tense in Background, Aim and Conclusion moves and past in Method and Results.

A similar analysis was performed by (Esfandiari, 2014) but regarding abstracts written in two subfields of Computer Science: Artificial intelligence and Architecture. The results, similar to those of (Tseng, 2011) show that authors prefer a four move structure and, in this case the use of the present form of verbs in all moves.

An extensive comparison between abstracts written in Persian and in English highlights cultural variations in (Zand-Vakili & Fard Kashani, 2012). The authors take into account the Information-Purpose-Methods-Products-Conclusion model of Hyland (2000) and the Create-A-Research-Space model of Swales (1990). The results show that both models have an intrinsic English base and that Persian articles present specific variations.

While all these studies point out strengths, weaknesses, and challenges related to abstract writing, a paper from 1958 written by Luhn presents the automatic creation of literature abstracts based on word frequency as a tool for indexing services and fast analysis of contents (Luhn, 1958). Since then a lot of work has been done on automatic summarizing with a lot of progress (Lloret & Palomar, 2012; Spärck Jones, 2007). With the emergence of nature inspired algorithms, a plethora of heuristics have also been used for text summarization (Araujo, 2007). Complex network techniques also have been employed with the same purpose (Antiqueira, Oliveira Jr., Costa, & Nunes, 2009). With such methods summarization evaluation techniques also have developed (Hariharan & Srinivasan, 2010; Owczarzak & Dang, 2009).

With specific emphasis on abstracting we can mention the automatic summarization system COMPENDIUM presented in (Lloret, Romá-Ferri, & Palomar, 2013). COMPENDIUM is used to generate abstracts from biomedical papers by using two approaches: one that selects the most relevant sentences from the document and one that is oriented towards actually generating the abstract of the paper. The results were evaluated both qualitatively and quantitatively showing the potential of the approach.

In between the two totally different abstract analysis approaches presented here, one that presents the linguists point of view regarding the structure and content and the other one, computer based, that aims at generating automatic abstracts based on the actual content of the text, our approach is situated in the middle by using computer science developed techniques to analyze the text and linguistic based recommendation to analyze results.

3. Method

The analysis is based on the Systemic Functional Linguistics register identification and variation of Halliday (2004) as used by Teich and Frankhauser (2010). The analyzed abstracts were extracted from a local journal (LJ), indexed in several scientific databases and recognized by the Romanian Executive Unit for Funding Education Higher, Research Development and Innovation, from 100 articles published from 2010 to 2014; and for comparison, from a top-ranked journal from the category Computer Science – theory and methods of the ISI Journal Citation Reports, as well 100 abstracts.

The ISI Journal Citation Reports was selected because it is currently the sole ranking system used for evaluating and promoting researchers in Romania. The top ranked journal (TRJ) was randomly selected from all the journals with impact factor above the median from the Computer Science – theory and methods category.

The abstracts were analyzed using the AntConc software (Anthony 2014) and tagged with parts of speech with the Stanford NLP tagger (Toutanova, Klein & Manning 2003, Toutanova & Manning 2000). The following five possible indicators of registry variation were computed:

i. The relative number of nouns (NN): scientific texts are expected to present a higher than average number of nouns;
ii. The relative number of lexical verbs (VV): a lower number of verbs is expected from a scientific text;
iii. The relative number of adverbs (ADV): also a lower number is expected;
iv. The type-token ration (TTR) which may indicate technical language. TTR is computed as the ration between number of the number of individual distinct words in a text (types) and total number of words in the text (tokens).
v. The lexical density (LEX) as measure of text density computed as the number of lexical word tokens (nouns, adjectives, verbs, adverbs) divided by the number of all tokens in the text.

4. Results and Discussion

Table 1 presents the total number of nouns, verbs and adverbs encountered in the selected abstracts, and the values of the type-token ratios and lexical densities. While these values indicate similarities between the two journals in terms of register variation in abstracts, further results obtained by analyzing abstracts separately (Table 2) the differences are significant.

Table 1. The five indicators computed for the three datasets. Values that potentially indicate best scientific writing for the given indicator are marked in boldface.

Indicator	LJ	TRJ
NN (%)	3694 (34.11%)	5445(34.19%)
VV (%)	295 (2.72%)	423 (2.65%)
ADV (%)	1753 (16.18%)	2450 (15.38%)
TTR	4.57	4.92
LEX	62.79	64.15

Table 2. Descriptive statistics regarding number of nouns (NN), adverbs (ADV) and verbs (VV), with Wilcoxon p-values indicating significant differences between them.

POS	Source	N Obs	Mean	Median	Lower 95% CL for Mean	Upper 95% CL for Mean	Wilcoxon p-value
ADV	LJ	90	0.46	0.42	0.39	0.53	0.0029
	TRJ	95	0.62	0.56	0.54	0.71	
NN	LJ	100	0.41	0.40	0.38	0.44	<.0001
	TRJ	100	0.60	0.51	0.55	0.66	
VV	LJ	100	0.42	0.43	0.39	0.45	<.0001
	TRJ	100	0.59	0.52	0.54	0.64	

The results presented in Table 2 show that abstracts from the local journal use significantly less adverbs than those form TRJ, with 10 of them actually adverb “free” and 5 of them from TRJ.

The same statistical difference is also present with regard to verbs, indicating that authors publishing in the local journal are aware of the good practices regarding the use of verbs, as well as adverbs, in scientific writing.

The situation differs however when it comes to nouns (NN): the significant difference shows that the number of nouns used in TRJ abstracts is higher (60% versus 41%) than those from LJ. Since nouns are used to represent concepts and terms, it follows that this difference may account for differences in writing that are only evident to researchers in the field.

Moreover, the Author's instructions in the LJ page encourage them to write the abstracts as short and concise as possible, limiting them to 250 words (this limitation does not apply to TRJ). It follows that authors – out of which 84% are Romanians – are summarizing their work as much as possible (the low count of adverbs and verbs supports this argument); however, in this "optimization" process they also reduce the number of nouns. This suggests that the abstracts are not necessarily concentrated in terms of information, but simply in terms of length.

5. Conclusions

While this simple analysis is not an evaluation of the quality of the articles published in the two journals, it could be used to derive some empirical conclusions regarding the writing habits of Romanian computer scientists.

The main result of such an analysis is that it can indicate some specific features to be targeted within an academic writing program. In this case, considering that authors publishing in the local journal are mostly Romanian researchers, it is worthwhile to include such a local analysis in the design of an academic writing program.

Further work consists in extending the study to other fields as well as to different but common parts of the scientific text (introduction, literature review, conclusions). A deeper analysis of abstract moves and possible correlation between different parts of speech and different moves can also offer interesting information about possible issues to be targeted into a scientific writing development program.

6. Acknowledgments

This work was possible with the financial support of the Executive Unit for Financing Higher Education, Research, Development and Innovation through the project Open School for Academic Self-improvement. Research, Academic Writing and Career Management (PN-II-PC-CA-2011-3.1-0682 212/2.07.2012)."

References

Anthony, L. (2014). AntConc (Version 3.4.3)[Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Antiqueira, L., Oliveira Jr., O. N., Costa, L. da F., & Nunes, M. das G. V. (2009). A Complex Network Approach to Text Summarization. *Inf. Sci.*, *179*(5), 584–599. doi:10.1016/j.ins.2008.10.032

Araujo, L. (2007). How Evolutionary Algorithms Are Applied to Statistical Natural Language Processing. *Artif. Intell. Rev.*, *28*(4), 275–303. doi:10.1007/s10462-009-9104-y

Biber, D. (1993). Using Register-diversified Corpora for General Language Studies. *Comput. Linguist.*, *19*(2), 219–241. Retrieved from http://dl.acm. org/ citation.cfm?id=972470.972472

Biber, D. (2006). *University language : a corpus-based study of spoken and written registers*. Amsterdam; Philadelphia: J. Benjamins.

Conrad, S. M. (1996). Investigating Academic Texts with Corpus-Based Techniques: An Example from Biology. *Linguistics and Education*, *8*, 299–326. doi:10.1016/S0898-5898(96)90025-X

Cross, C., & Oppenheim, C. (2006). A genre analysis of scientific abstracts. *Journal of Documentation*, *62*(4), 428–446. doi:10.1108/00220410610700953

Esfandiari, R. (2014). Realization of Rhetorical Moves and Verb Tense Variation in Two Subdisciplines of Computer Sciences : Artificial Intelligence and Architecture. *International Journal of Language Learning and Applied Linguistics World*, *5*(February), 564–573.

Halliday, M. A. K. (2004). *An introduction to functional grammar*. *Lingua* (Vol. 3rd ed /, p. 689). doi:10.1016/0024-3841(86)90084-7

Hartley, J. (1994). Three ways to improve the clarity of journal abstracts. *British Journal of Educational Psychology*, *64*(2), 331–343. doi:10.1111/j.2044-8279.1994.tb01106.x

Hariharan, S., & Srinivasan, R. (2010). Studies on Intrinsic Summary Evaluation. *Int. J. Artif. Intell. Soft Comput.*, *2*(1/2), 58–76. doi:10.1504/IJAISC.2010.032513

Hyland, K. (2000). *Disciplinary discourses : social interactions in academic writing*. Harlow, England; New York: Longman.

Kaplan, R., Cantor, S., Hagstrom, C., et al. (2009). On abstract writing. *Text - Interdisciplinary Journal for the Study of Discourse*, 14(3), pp. 317-453. Retrieved 29 Nov. 2014, from doi:10.1515/text.1.1994.14.3.401

Lloret, E., & Palomar, M. (2012). Text Summarisation in Progress: A Literature Review. *Artif. Intell. Rev.*, *37*(1), 1–41. doi:10.1007/s10462-011-9216-z

Lloret, E., Romá-Ferri, M. T., & Palomar, M. (2013). Editorial: COMPENDIUM: A Text Summarization System for Generating Abstracts of Research Papers. *Data Knowl. Eng.*, *88*, 164–175. doi:10.1016/j.datak.2013.08.005

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*. doi:10.1147/rd.22.0159

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *Proceedings of the 21st National Conference on Artificial Intelligence*, *1*, 775–780. doi:10.1.1.65.3690

Owczarzak, K., & Dang, H. T. (2009). Evaluation of Automatic Summaries: Metrics Under Varying Data Conditions. In *Proceedings of the*

2009 Workshop on Language Generation and Summarisation (pp. 23–30). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1708155.1708161

Santos, M. (2009). The textual organization of research paper abstracts in applied linguistics. *Text - Interdisciplinary Journal for the Study of Discourse*, 16(4), pp. 449-588. Retrieved 29 Nov. 2014, from doi:10.1515/text.1.1996.16.4.481

Spärck Jones, K. (2007). Automatic Summarising: The State of the Art. *Inf. Process. Manage.*, *43*(6), 1449–1481.doi:10.1016/j.ipm. 2007.03.009

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

Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills*. Ann Arbor, Mich: University of Michigan Press.

Teich, E., & Fankhauser, P. (2010). Exploring a corpus of scientific texts using data mining. *Corpus-Linguistic Applications: Current Studies, New Directions*, 233–248.

Tibbo, H. R. (1992). Abstracting across the Disciplines: A Content Analysis of Abstracts from the Natural Sciences, the Social Sciences, and the Humanities with Implications for Abstracting Standards and Online Information Retrieval. *Library and Information Science Research*, *14*, 31–56.

Toutanova, K., Klein, D., & Manning, C. D. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1 (NAACL '03),* (pp. 252–259). doi:10.3115/1073445.1073478

Toutanova, K., & Manning, C. D. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (pp. 63–70). doi:10.3115/1117794.1117802

Tseng, F. (2011). Analyses of Move Structure and Verb Tense of Research Article Abstracts in Applied Linguistics. *International Journal of English Linguistics*, *1*(2), 27–39. doi:10.5539/ijel.v1n2p27

Zand-Vakili, E., & Fard Kashani, A. (2012). The Contrastive Move Analysis: An Investigation of Persian and English Research Articles' Abstract and Introduction Parts. Mediterranean Journal of Social Sciences, 3(2), 129–137. doi:10.5901/mjss.2012.v2n3