Approaches for Helping Brazilian Students Improve their Scientific Writings

Ethel Schuster[1], Rick Lizotte[1], Sandra M. Aluísio[2], Carmen Dayrell[3]

[1]Northern Essex Community College – 100 Elliott Street – Haverhill, MA 01830, USA

2NILC/ICMC, USP – CP 668 13560-970 – São Carlos – SP – Brazil

[3]UNINOVE – Av. Dr. Adolpho Pinto, 109 – 01156-050 - São Paulo – SP – Brazil

{eschuster,rlizotte}@necc.mass.edu,sandra@icmc.usp.br, dayrellc@gmail.com

Abstract. *Writing well in English is a challenge for non-native English speakers. When readers are unable to comprehend what they read they just give up reading and fail to get to the content. In this paper we describe problems encountered in the English writing of scientific abstracts by Brazilian Portuguese speakers. We collected and analyzed a corpus of 115 abstracts in which we identified specific language-related errors that affect comprehension. We show that students who must write scientific papers may benefit significantly with practice exercises, and computer-based tools. The use of such tools can enhance the students' level of confidence and thus enable them to improve their writing.*

1. Introduction

Academic researchers anywhere in the world must publish in English. However, their ability to produce well-written documents that get published is hindered by their linguistic capabilities. Researchers at Núcleo Interinstitucional de Linguística Computacional (NILC) have studied scientific writing over the last 20 years [Fontana et al. 1993; Aluísio and Oliveira 1995; Aluísio et al. 2001; 2005; Schuster et al. 2005; Dayrell et al. 2012]. They have developed techniques and software tools to assist novice researchers with their writing. Still, many students fail to publish their research findings because the grammatical and lexical errors in their writing interfere with their ability to convey their message clearly. Writing well is a pervasive problem among non-native speakers of English and several researchers [Han et al. 2006; Genoves et al. 2007; Lee 2009; Umezawa et al. 2013] have focused on this problem. This paper describes specific problem areas that can help students improve their writing. Section 2 describes our corpus-based approach. Section 3 discusses what we have learned from this analysis; we can turn these into teaching strategies to help writers.

2. Methods and materials: A Corpus-based Approach

We collected 115 abstracts from students enrolled in five graduate scientific writing courses at several universities in Brazil, beginning in 2004[1]. Students came from various disciplines, including pharmacy, chemistry, biology/genetics, physics, and computer

[1] This corpus is available if asked for. Names of students were not included in order to preserve anonymity.

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 178–182,
Fortaleza, CE, Brazil, October 21–23, 2013.

science. As part of the course's requirements, the students had to write a research paper. The abstracts were collected from these final papers.

2.1 Tagging Errors

A native English speaker who is a linguist and teaches English as a Second Language tagged 23 error categories in each abstract. The errors were organized into three groups: (1) mechanical, e.g., the use of punctuation (tagged as P), capitalization (CAP), and spelling (SP); (2) lexical; and (3) syntactic. Lexical use included word use errors (WU), the incorrect use of a word to express an intended meaning, such as **"**Mutants were **availed**" instead of **"**Mutants were **used**"; word use collocation errors (WUCol), the incorrect use of lexical items in idioms and common collocations, such as *depend **of*** instead of *depend **on***; and word form errors (WF), the incorrect use of common word forms, such as *this/these, that/those*. Categories for syntactic accuracy included those dealing with article use (ART-, ART+, ART), word order (WO_NP, WO_ADJ, WO_S, WO), Subject-Verb-Object structure of the clause (S+VO, S-VO, SV-O), part-of-speech (POS), verb use (VU, VF, SVA), and the use of singular/plural in nouns and adjectives (S/PL, S/PL_ADJ). At times, abstract writers incorrectly used a Portuguese word or the Portuguese spelling of a technical term (PORT). These 23 categories comprehensively covered all the errors we found. Six error categories (WU, ART-, P, SP, WUCol, ART+) accounted for 66% of all errors. The distribution (Rank, Error, Count, Percent) was: 1, WU, 497, 25.8%; 2, ART-, 258, 13.4%; 3, P, 165, 8.6%; 4, SP, 147, 7.6%; 5, WUCol, 109, 5.7%; 6, ART+, 95, 4.9%. Total: 1271 errors, 66.0%.

3. Discussion: What can the results teach us?

Here, we focus on errors related to lexical use (WU and WUCol) and those involving the incorrect use of articles (ART- and ART+). We discuss our results and suggest ways to correct writers' errors.

3.1 Word-Use errors (WU)

These errors result when an incorrect word is used. For example, let us compare the use of **"amount"** and **"number"** when describing a sequence of multiple steps called "alignments." Since the number of steps is countable, the correct phrase would be "The **number** of alignments." However, very often we encounter the incorrect use of "*The **amount** of alignments."[2] On the other hand, the correct form, "the **amount** of information" (which cannot be counted) is often confused by non-native English speakers with the twice incorrect countable form, "*the **number** of **informations**." Another common error is "*many researches," which is a direct (but incorrect) translation from the Portuguese equivalent of "many papers." Research cannot be counted, but papers can be.

3.2 Word Use-collocation errors (WUCol)

Collocations are strings of words that go together. Many times, students use wrong word combinations. For example, they may use the wrong preposition such as: "* The search **of** genes responsible…" rather than "The search **for** genes responsible…"

---

[2] Here we use the "*" as standard convention to mark an ungrammatical sentence.

3.3 Article errors (ART-, ART+)

We have identified sentences in which the lack of an article (the words "the", "a") (ART-) interferes with the ability of the reader to understand the text and thus the content of the paper. For example: "* Pollen is ∅ male gamete of higher plants…" instead of "Pollen is **the** male gamete of higher plants…" Our writers, however, generally tended to include an article where it should not be present (ART+). One Portuguese writer included a superfluous article, which introduced ambiguity in the text: "In this paper, we write about the protein. We discuss how the bacteria grows and we discover how the protein has …" If specific protein or bacteria have not been mentioned beforehand, one correct form would be: "In this paper, we write about proteins. We discuss how bacteria grow and we discover how the proteins have …" The definite and indefinite use of articles depend on several factors: (1) countability of the modified noun, (2) generality of the referent, (3) definiteness of the referent, and (4) possible previous mention of the referent.

3.3.1 Countable and General

If the noun is countable and meant to refer to a general category, the most common choice is the plural noun without any article. For example, instead of "***The** computer is equipment that can improve **the** math learning of ***the** students." (first sentence in abstract), the correct sentence would not include the articles: "Computers are pieces of equipment that can improve math learning for students." Computers and students are described in general but none of them have been specifically referred to before.

3.3.2 Uncountable and General

If the noun is not countable and meant to refer to a general description, the most common choice is the singular noun without any article. For example, "Alcohol electrooxidation is a theme … studied in electrocatalysis." does not include the article as in "***The** alcohol electrooxidation is a theme … studied in **the** electrocatalysis." The sentence refers to a process in general and not to a particular instance of its use. One explanation for finding so many (ART+) errors is that Romance languages tend to use the definite article in this case in which English does not.

3.3.3 Countable/Uncountable, Specific and Indefinite

If there is a specific but indefinite reference for a noun and the noun is countable, the most common choice is the article "a"/"an". For instance, **"**This work shows the necessity for **a** *hole injection layer*." is the correct sentence instead of "*This work shows the necessity for ∅ hole injection layer." There is a specific instance of "a hole injection layer" mentioned for the first time, but it is not referring back to a particular one. If the noun is uncountable, no article is used, as in "Our results show the necessity for ∅ further research on this topic."

3.3.4 Countable/Uncountable, Specific and Definite: Anaphoric Use

If specific reference to an instance(s) of a noun has been made using an indefinite article *a/an/plural* for countable nouns or *nothing* for uncountable nouns, the referent of that noun can be "pointed to" using the definite article *the*. For example (uncountable)**: "**This work uses egg white …to verify ... **The** egg white (anaphoric use) underwent two heating

rates." instead of "This work uses egg white …to verify ... * ∅ Egg white underwent two heating rates."

This example (countable) includes the article, "A total of ... *tests* were conducted…The results of **the** tests indicated that …" compared to the original with the missing article "*A total of … *tests* were conducted…The results of ∅ tests indicated that …" This is called the "second mention" use of the definite article, or anaphoric use of the definite article. It can be explained as *pointing backwards* to the indefinite use of the noun.

3.3.5 Extension of Anaphoric Use of Definite Article

If the referent of a noun has been introduced into the discourse, any part or process associated with the referent can be pointed to with the definite article, as in "We investigated the … using *an experiment that*…**The** results demonstrated …"

3.3.6 Cataphoric Use

When both the referent and its associated part or process can be introduced within the same noun phrase (NP), this is called a cataphoric use. The definite noun indicating the part or process can occur *before* the noun that introduces the whole into the discourse. For example, an abstract could begin with "The results of *an experiment* testing … demonstrated that …" where the results are part of the experiment, first mentioned in *an experiment*. The article "**the**" in "the results" points forward to *an experiment*. The lack of cataphoric reference use with definite articles was so noticeable in our data that we find it necessary to teach the specifics of their use.

4. Concluding Remarks

We have identified errors that can be taught to improve writing quality and readability. Computer-based tools are very useful. Errors related to lexical use (WU and WUCol) can be reduced by consulting the Corpus of Contemporary American English (COCA)[3]. For example, no matching records are found for "*the **number** of **information**". Students could search for a noun to replace "number" while keeping the intended meaning. A total of 171 instances are displayed for "the **amount** of information"; clearly, "amount" is the best choice. COCA can also help choose the right preposition. For example, we found 31 instances of "the search **of**". In most cases, "of" is followed by a noun that indicates either data (major databases, the literature, records), location (the house, his computer, student luggage) or the agent performing the action (the Federal Police). Querying "what preposition follows 'the search'?", "for" yields 1,100 hits. The examples show that "the search **for**" is followed by a noun that refers to whatever you are looking for. Han [Han et al. 2006] and Genoves [Genoves et al. 2007] implemented machine-learning systems that automatically detect English article usage errors. Unfortunately, these are not yet available to users. We hope to make them available within our writing tools soon. We also plan to assess quality improvements in students' writing after their using corpus-based tools such as COCA.

[3] COCA is a 450-million word corpus and includes academic texts, fiction, spoken language, popular magazines, and newspapers. It is freely available from http://corpus.byu.edu/coca/.

References

Aluísio, S. M., Schuster, E., Feltrim, V. D., Pessoa Jr., A., Oliveira Jr., O. N. (2005) "Evaluating Scientific Abstracts with a Genre-specific Rubric", In: 12th International Conference on Artificial Intelligence in Education (AIED 2005), 2005, Amsterdam, 2005. v. 1. pp. 738-740.

Aluísio, S. M., Barcelos, I., Sampaio, J., Oliveira Jr, O. N. (2001) "How to Learn the Many Unwritten ´Rules of the Game´ of the Academic Discourse: A Hybrid Approach Based on Critiques and Cases to Support Scientific Writing", In: IEEE INTERNATIONAL CONFERENCE ON ADVANCED LEARNING TECHNOLOGIES, 2001, Madison, Wisconsin. Los Alamitos, CA: IEEE Computer Society, 2001. v. 1, pp. 257-260.

Aluísio, S. M. and Oliveira Junior, O. N. (1995) "A Case-Based Approach for Developing Writing Tools Aimed at Non-native English Users", In the Proceedings of The 1st International Conference, ICCBR-95, Sesimbra, Portugal, pp. 121-132.

Dayrell, C., Candido Jr. A., Lima, G., Machado Jr. D., Copestake, A. A., Feltrim, V.D., Tagnin, S.O. and Aluísio, S.M. (2012) "Rhetorical Move Detection in English Abstracts: Multi-label Sentence Classifiers and their Annotated Corpora", LREC 2012, pp. 1604-1609.

Fontana, N., (Caldeira), Aluísio, S.M., De Oliveira, M.C.F. and Oliveira Jr., O.N. (1993) Computer Assisted Writing Applications to English as a Foreign Language. CALL, Volume 6 (2), pp. 145-161.

Genoves Jr., L.C., Lizotte, R., Schuster, E., Dayrell, C., Aluísio, S.M. (2007) "A two-tiered approach to detecting English article usage: an application in scientific paper writing tools", In the Proceedings of The 6th International Conference Recent Advances in Natural Language Processing (RANLP 2007), Borovets, Bulgaria, pp. 225-229.

Han, N., Chodorow, M. and Leacock, C. (2006) Detecting errors in English article usage by non-native speakers. Nat. Lang. Eng. 12, 2 (June 2006), pp. 115-129.

Lee, J. S. Y. Automatic Correction of Grammatical Errors in Non-native English Text. (2009) Ph. D. Thesis. Massachusetts Institute of Technology, 2009, http://groups.csail.mit.edu/sls/publications/2009/Thesis_Lee.pdf

Schuster, E., Aluísio, S. M., Feltrim, V. D., Pessoa Jr, A., Oliveira Jr., Osvaldo N. (2005) "Enhancing the Writing of Scientific Abstracts: A Two-phased Process Using Software Tools and Human Evaluation", In: Encontro Nacional de Inteligência Artificial (ENIA 2005), São Lourenço, Brazil, pp. 962-971.

Umezawa, J., Mizuno, J., Okazaki, N., Inui, K. (2013) "Evidence in automatic error correction improves learners' English skill", Proceedings of the 14th international conference on Computational Linguistics and Intelligent Text Processing (CICLing 2013). Berlin: Springer-Verlag, 2013, pp. 559-571.