English for Specific Purposes 39 (2015) 26–44

Contents lists available at ScienceDirect

English for Specific Purposes

journal homepage: http://ees.elsevier.com/esp/default.asp

The vocabulary of agriculture semi-popularization articles in English: A corpus-based study

Verónica L. Muñoz*

Facultad de Ciencias Humanas – Facultad de Ingeniería, Universidad Nacional de Río Cuarto, Ruta 36 – Km 601, Argentina

ARTICLE INFO

Article history:
Available online

Keywords:
Semi-popularization article
Technical and non-technical vocabulary
Agriculture

ABSTRACT

Scientific communication can be represented as a continuum with the research article, addressed to specialized researchers, at one end of the spectrum, and with the popularization article, addressed to lay readers, at the other. In between there exists a distinct genre, the semi-popularization article, which has not received much attention in the English literature. As a contribution to the field, this paper describes the vocabulary of a corpus of 700 agriculture semi-popularization articles in English. The analysis was carried out in two stages that combined quantitative and qualitative methodology: a lexical description of the corpus and the analysis of high-frequency words. The lexical analysis revealed high lexical variation in the corpus and narrow word range. Academic words provided a lower coverage (6%) than that usually reported for research articles (10–12%), and a higher coverage than that reported for newspapers (4%). The analysis of high-frequency words showed that many of these words, including general and academic words, were closely associated with the discipline of agriculture, and therefore represented the technical vocabulary of the texts.

© 2015 Published by Elsevier Ltd.

1. Introduction

Scientific English, as described by Halliday (2004), is not homogenous but varies in relation to the contexts where the language is used. Such contexts range from clearly defined discourse communities, the domain of science, to more general and heterogenous groups of individuals, the domain of the public. Within these contexts, individuals engage in communication by participating in the production and reception of texts which vary in their degree of specialization. Viewed in this way, scientific communication can be represented as a continuum with the research article, addressed to specialized researchers, at one end, and with the popularization article, addressed to a lay readership, at the other (Ciapuscio, 2003).

Different studies have explored linguistic features that characterize research articles and popularization articles. One aspect that has gained particular attention is vocabulary, recognized as one of the most distinctive features of scientific discourse (Bowker & Pearson, 2002; Cabré, 1999; Ciapuscio & Kuguel, 2002; Halliday, 1993, 1998, 2004). Drawing on corpus data, many studies have provided insightful evidence about the different types of vocabulary used in research articles. Most studies have focused on the analysis of academic words in research articles from different disciplines, such as medicine (Chen & Ge, 2007; Wang, Liang, & Ge, 2008), agriculture (Martínez, Beck, & Panza, 2009), chemistry (Valipouri & Nassaji, 2013), and applied linguistics (Khani & Tazik, 2013). Other studies have examined the use of academic and non-academic words in

* Lamadrid 115 3 C, CP 5800, Río Cuarto, Córdoba, Argentina. Tel.: +54 358 4624142.
E-mail address: vmunoz@hum.unrc.edu.ar.

http://dx.doi.org/10.1016/j.esp.2015.04.001
0889-4906/© 2015 Published by Elsevier Ltd.

applied linguistics research articles (Vongpumivitch, Huang, & Chang, 2009). A number of studies have compared the use of academic words in research articles and university textbooks (Coxhead, 2000), textbook chapters, academic book reviews, master's theses, and doctoral dissertations (Hyland & Tse, 2007), and textbook chapters, science squibs, and academic book reviews (Hyland & Tse, 2009). These studies have made contributions to the field by producing discipline-specific wordlists.

Vocabulary has been also analyzed in popularization articles appearing in different publication outlets with varying levels of specialization, such as newspapers (Adams Smith, 1987; Gallardo, 1998; Mapelli, 2004; Myers, 1990, 1994), general interest magazines (Adams Smith, 1987), and specialized magazines (Mayor Serrano, 2003; Mapelli, 2004; Myers, 1990, 1994). These studies have largely focused on the lexical resources used to reduce the degree of technicality of popularization articles and thus facilitate the non-specialist reader's comprehension of terminology. However, to the best of my knowledge, there is no published study which has sought to provide a categorization of the types of vocabulary used in popularization articles. More importantly, the studies analyzing popularization articles have explored texts from a variety of publication formats that have varying levels of specialization and address different readers, overlooking differences that characterize the various contexts where science is popularized, such as the mass media, universities, and schools. In fact, few studies, especially in English, have described one particular type of popular science text, the semi-popularization article, which has been identified as a distinct genre with characteristics that differ from those that define the popularization article (Ciapuscio & Kuguel, 2002; Gallardo, 1998), the most important being the writers, the audience, the publication format, and the communicative purpose. Semi-popularization articles are generally written by researchers, whereas popularization articles are usually written by journalists. Semi-popularization articles are addressed to readers with some expertise and knowledge in a discipline, whereas popularization articles are addressed to the general public. Semi-popularization articles are published in magazines with some level of specialization, whereas popularization articles are published in outlets such as newspapers. As for the communicative purpose, semi-popularization articles inform readers about scientific findings, tools, and developments to people in academic and research institutions and in some spheres of society, whereas popularization articles present research findings that are novel to society and persuade the general public about the potential consequences of science in their daily lives (Alcíbar, 2004; Ciapuscio, 1997).

The contextual features that distinguish semi-popularization articles from research articles and from popularization articles make this genre particularly suitable for pedagogic purposes in ESP reading courses at undergraduate level. As in many Latin American countries (Dudley-Evans & St. John, 1998), in our university, Universidad Nacional de Río Cuarto (UNRC), Argentina, undergraduate ESP courses are aimed at preparing students to read in one specialized field for professional and academic purposes. The reading materials used in these courses usually include authentic texts related to specific disciplines. With the exception of the hard sciences, in which students are trained to read academic texts such as research articles very early in their career programs, most ESP courses at UNRC generally use texts such as semi-popularization articles. These texts are less specialized and less conceptually complex than research articles, as they are addressed to readers who do not have much expertise in particular areas of knowledge, such as undergraduate students who are being trained in a specific discipline and who do not usually have enough training in reading scientific genres. Being targeted to this type of audience, semi-popularization articles do not require much familiarity and expertise in a specific discipline, as opposed to research articles, which generally "tend to be long and demanding in their content and in the language used to express that content" (Coxhead & Byrd, 2007, p. 133). In sum, semi-popularization articles are more accessible to students who have no experience in reading technical texts and can therefore help bridge the gap between general English and more specialized texts in ESP courses.

Given the pedagogic value in using semi-popularization articles in undergraduate ESP reading courses, the study of the types of vocabulary used in this genre can inform teaching practices. The findings can bring to light lexical features of semi-popularization articles so that ESP teachers can be better prepared to use this genre for the development of reading skills, both by raising students' awareness about the characteristics of the genre and by building wordlists specifically tailored for reading this genre in ESP contexts.

The purpose of this paper is to report on a corpus study about the vocabulary of agriculture semi-popularization articles in English. The first stage of analysis focused on a general lexical description of the corpus, particularly on vocabulary size, standardized type/token ratio, and word range, as well as the coverage of grammar words, general words, and academic words. The second stage focused on the analysis of the high-frequency words in the corpus.

2. Theoretical background

2.1. Scientific communication

Following Swales's (1990) definition of genre as texts with particular textual conventions created in response to routine social activities in specific cultural contexts and for specific purposes, this study adheres to the view that scientific communication can be described as a continuum featuring three main genres: the research article, the semi-popularization article, and the popularization article (Ciapuscio, 2003). At one end of the continuum, the research article represents the genre through which researchers communicate within specific academic discourse communities. Research articles are written by scientists and are addressed to other scientists in the same discipline. These articles are published in highly specialized refereed scientific journals that are usually issued periodically. Through the research article, researchers present their findings to the scientific community, engage in scholarly debate to discuss and evaluate their contributions to the field, and make claims in order to construct scientific knowledge. The research article is, then, "the principal site of disciplinary

knowledge-making" (Hyland, 2010, p. 117). At the other end of the continuum, the popularization article represents the dissemination of science to a heterogenous non-specialist audience in the public domain (Calsamiglia, 1997). Popularization articles are written by journalists and are addressed to lay readers who lack knowledge and expertise in specific disciplines. Popularization articles are published in the media (Calsamiglia, 2003), such as newspapers and general interest magazines that are generally issued weekly or daily. The main purpose of popularization articles is to communicate scientific findings to society and persuade a wide readership about the impact of science in their daily lives (Ciapuscio, 1997). In order to achieve these goals, popular science texts recontextualize scientific knowledge in the form of news or scientific breakthroughs (Hyland, 2010) by presenting research findings that are relevant to real-life concerns, and by emphasizing the social consequences of science (Alcíbar, 2004; Ciapuscio, 1997). In this process, journalists mediate between expert researchers and lay readers by connecting "the abstractions of scientific knowledge to lived experience" (Myers, 2003, p. 269).

2.1.1. The semi-popularization article

In between the research article and the popularization article, there exists a particular type of text which some authors (Ciapuscio & Kuguel, 2002; Gallardo, 1998) have identified as a distinct genre, the semi-popularization article. As stated by Ciapuscio and Kuguel (2002) and Gallardo (1998), semi-popularization articles are usually written by researchers, and may occasionally be written by popularizers who are trained for communicating scientific and technological information in specialized disciplines. These articles are addressed to readers who are not specialists but have some level of proficiency and expertise in a specific area of knowledge, described by Ciapuscio (2003) as "(semi-) laypersons" who have "an intermediate degree of competence lying somehow in between the extremes of […] the expert and the lay person" (p. 230), such as university students being trained in a particular discipline. These readers represent a wider audience than the highly specialized academic communities reading research articles, but constitute a more restricted readership than the general public reading popularization articles in newspapers. As for the publication format, semi-popularization articles are published in scientific popularization magazines, such as *American Scientist*, *New Scientist* and *Science*. The communicative purpose of these texts is to inform readers about scientific findings, technology, and developments to people in particular spheres of society. Figure 1 synthesizes the major differences between the research article, the semi-popularization article, and the popularization article.

2.2. English vocabulary

2.2.1. General words and academic words

The vocabulary of science in English has been generally described using two established wordlists in the literature: West's General Service List or GSL (1953) and Coxhead's Academic Word List or AWL (2000). The GSL contains general words typically defined as non-specialist words that occur frequently in a variety of language uses and texts (Brezina & Gablasova, 2013; Coxhead & Nation, 2001; Nation, 2001b). These words include approximately 2,000 word families containing content words and most of the grammar words in English. General words usually cover about 80 percent of the running words in a text. Despite its wide use, the list has been criticized, two of the main problems identified being the fact that it is out of date and that it was built based on word families. In response to these criticisms, a New General Service List (new-GSL) has recently been constructed by Brezina and Gablasova (2013), who used four corpora representing different time periods, lemmas as a

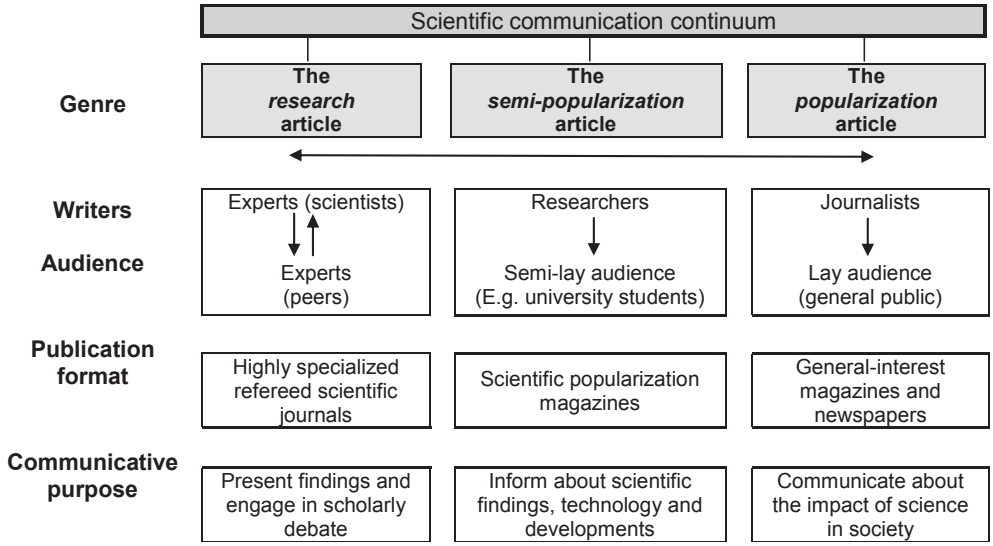

Figure 1. Major differences between the research article, the semi-popularization article and the popularization article.

unit of analysis, and frequency, dispersion, and distribution criteria. The new-GSL contains 2,494 lemmas which cover about 80–81 percent of texts.

The AWL (Coxhead, 2000) comprises academic words which are not general but occur frequently over a wide range of academic texts from different disciplines (Nation, 2001a). Academic vocabulary usually covers about 8.5–10 percent of the running words in academic texts, 3.9 percent of newspapers, and 1.7 percent of fiction texts. Academic words and general words together cover about 90 percent of the running words in an academic text (Coxhead, 2000; Coxhead & Nation, 2001; Nation, 2001b). Like West's GSL, Coxhead's AWL has been criticized, largely because the list was built excluding the words from the GSL and using word families. In response to these criticisms, the Academic Vocabulary List (AVL) has been recently constructed by Gardner and Davies (2013), who used a 120-million-word corpus representing nine academic disciplines, lemmas as unit of analysis, and ratio, range, dispersion, and discipline measure criteria. The AVL contains 3,000 lemmas and the top 570 word families cover about 14 percent of academic texts.

2.2.2. Non-technical words and technical words

Two other categories, particularly relevant for this study, are non-technical and technical words. The notion of non-technical words is especially interesting because, unlike the academic words represented in the AWL and the AVL, they are conceptualized in terms of semantic and pragmatic criteria rather than quantitative criteria. As P. Meyer (1997) claims, non-technical vocabulary builds the semantic–pragmatic skeleton of academic texts, as it determines the status and the relations of the subject-specific propositions. As defined by P. Meyer, non-technical words are used to express semantic and pragmatic meanings that are not strictly discipline-oriented and usually refer to entities, properties, and processes that are not specific to one particular discipline, such as *problem*, *theory*, *considerable*, *importance*, *confirm,* and *findings*. As non-technical words "are semantically closely connected with the process of academic research and communication, most of them denoting activities, achievements, accomplishments, and mental states in the process of scholarly investigation" (P. Meyer, 1997, p. 7), their reference is more general than technical words and can be associated with a wide range of specialized contexts. Non-technical words account for an important proportion of the vocabulary in specialized texts.

Technical words are discussed in the literature from different perspectives, two of which are central for the purpose of this paper: the applied linguistics perspective (Chung & Nation, 2004; Coxhead & Nation, 2001; Nation, 2001a, 2001b) and the perspective of terminology (Bowker & Pearson, 2002; Cabré, 1999; Pearson, 1998). As Nation (2001a) observes, there has been little research which provides "a consistently applied operational definition of what words are technical words" (p. 18). In the field of applied linguistics, Chung and Nation (2004) define technical words as those that are very frequent in a specific text and a specialized domain but are infrequent or non-existent in other fields. Although it is not possible to calculate exactly the total amount of technical words in a discipline, Coxhead and Nation (2001) suggest that the technical vocabulary of a discipline accounts for probably 1,000 words or less, which is the typical amount of vocabulary included in technical dictionaries. Instances of technical words are acronyms, abbreviations, chemical formulas, symbols, and those words associated with a particular field (Nation, 2001a). From the perspective of terminology, a distinction is made between terms and words. Terms, or technical words, are defined as distinctive meaningful units of an established code in specialized languages "used to designate concepts pertaining to special disciplines and activities" (Cabré, 1999, p. 81). Unlike terms, which have special reference, words are items with general reference (Pearson, 1998).

Different methods are discussed in the literature for identifying technical words in a corpus, a task that cannot always be done straightforwardly (Chung & Nation, 2004; Pearson, 1998). As stated by Cabré (1999), Bowker and Pearson (2002), and Chung and Nation (2003), some of these procedures draw on statistical formulas and quantitative data, such as comparing the frequency and range of words in a specialized corpus and in a general corpus or in another discipline, and using term extraction procedures. However, frequency criteria can be misleading in that the procedure may overlook the actual meanings that words acquire in different contexts. As observed by Pearson (1998), words used infrequently in everyday language may have one meaning in the general language and a different meaning in specialized communicative settings. In addition, as Cabré (1999) and Coxhead and Nation (2001) point out, frequency counts typically reveal that many of the topic-related words in a corpus are actually general words that acquire a more specialized meaning in particular fields, or words from certain disciplines that are used in other fields. What these claims bring to light is that technical words include not only highly specialized words which have a single meaning and occur frequently in a discipline but also words that are formally like general words but have a specialized meaning in a particular field or words that are used in different disciplines with different meanings. From this perspective, examples of technical words are *homonymy* and *meaning* in linguistics, *miogeoclinal* and *arcs* in geology, *flavonoids* and *bacteria* in microbiology, *spectrophotometer* and *currents* in engineering, and *mesothelioma* and *culture* in medicine.

Two quantitative approaches used for the identification of discipline-specific words are, as specified by Coxhead (2011), selecting vocabulary using the GSL and the AWL as base lists, as in Coxhead and Hirsh (2007), and selecting vocabulary without using the lists, as in Ward (2009). The downside of the first approach is that it presupposes definite boundaries based on the assumption that what is general in the GSL and what is academic in the AWL is not technical, thus conceptualizing specialized vocabulary as that which occurs outside the GSL and the AWL. However, these conceptual divisions are not universal and common to all disciplines because, as explained above and as Schmitt and Schmitt (2012) point out, academic and technical vocabulary may be subsets of general English. By contrast, the value of the second approach is that it allows researchers to identify not only technical words outside the GSL and the AWL which have a unique specialized meaning but also words from the GSL and the AWL that acquire technical meanings in a specific discipline. However, this second approach

may better represent the uses that words have in scientific disciplines and genres if the selection of vocabulary is based on qualitative evidence, that is word meaning, in addition to frequency evidence.

One of the methods that relies on word meaning involves drawing on semantic and pragmatic criteria, as proposed in the field of terminology (Cabré, 1999). The strength of the method is that it helps to identify the actual meanings of words as used in specific genres and disciplines and to consider not only highly specialized words unique to a field but also words that acquire a specialized meaning in a corpus. The approach accounts for the ways meaning is determined by relations among words that belong to the lexicon of a discipline and also by the specialization of the contexts where words are used. From a semantic perspective, technical words operate as reference units representing concepts of a discipline, which may denote entities, processes, actions, and properties (Cabré, 1999). The close association between technical words and scientific concepts is fundamental, as these words represent meanings which are semantically linked in a system of inter-related concepts (Pearson, 1998). From this perspective, a defining condition for a word to be a technical term is that it acquires its specialized meaning as a member of a set of terms and concepts which belong to a specific field (Cabré, 1999; Pearson, 1998). From a pragmatic perspective, technical words are terms used in well-defined, specialized communicative situations (Cabré, 1999; Pearson, 1998), as in the case of scientific texts. As identified by Bowker and Pearson (2002), Cabré (1999) and Pearson (1998), the contexts where science is communicated are largely shaped by the level of expertise and disciplinary knowledge of those who participate in the process of communication. For instance, in some contexts communication takes place among experts who have the same level of specialized training and experience in a particular discipline, as in the case of research articles. These situations require highly specialized vocabulary shared by writers and readers. There exist other contexts, as in the case of semi-popularization articles, in which communication generally takes place between experts and semi-experts, who have some level of disciplinary knowledge but do not have the experts' level of expertise, such as university students. Given this asymmetry in the writers' and readers' level of expertise, the amount and specialization of words is usually reduced, with some technical terms usually explained through the use of general language. Other contexts where the communication of science may take place are those situations where no previous disciplinary knowledge is assumed from the target audience, as in the case of popularization articles published in the press.

2.3. Research questions

Using not only quantitative methods but also qualitative procedures, this study addresses the following research questions:

1. For the general lexical description of the corpus:
   1.1. What is the vocabulary size of the corpus?
   1.2. What is the range of words?
   1.3. What is the coverage of grammar words, general words, and academic words?
2. For the analysis of high-frequency words in the corpus:
   2.1. Which are the high-frequency words of the corpus?
   2.2. Is the classification of high-frequency words based on two wordlists available in the literature, the GSL and the AWL, truly representative of the meanings of words as used in agriculture semi-popularization articles?
   2.3. How can high-frequency words be categorized according to the meanings they acquire in agriculture semi-popularization articles?

Addressing these questions will shed light on the types of vocabulary used in agriculture semi-popularization articles and will enable the compilation of wordlists specifically tailored for teaching reading skills in undergraduate ESP courses, where semi-popularization articles are particulary suitable for pedagogic purposes.

3. Materials and methods

3.1. The corpus

A specialized corpus was built to represent a genre (semi-popularization articles), a register (science), a discipline (agriculture), a specific domain (corn production), and a language (English). To maximize representativeness, the corpus was designed based on criteria proposed by Sinclair (1991, 2005), C. Meyer (2004), and Biber (2008): balance, diversity of sources, availability of texts in electronic form, period, size, use of complete texts, and variety of writers. Balance was determined by the corpus' internal structure, as it was planned to be specific in terms of discipline, text domain, genre, register, and language. The decision to focus on one domain – corn production – was taken for two reasons. First, the area of corn production is in itself large enough to represent many of the subjects researched in agriculture. Second, as will be explained later, the analysis of high-frequency words was based on semantic and pragmatic criteria, which required close observation of the meanings words acquired in the corpus as members of a conceptual field, a task that was facilitated by focusing exclusively on one subarea of agriculture.

To account for diversity of sources and availability of texts in electronic form, the texts were retrieved from twelve popularization magazines published online by eleven American universities. The number of articles selected from each magazine depended on their publication frequencies, that is whether they were issued quarterly, bimonthly, monthly, or weekly (Appendix A), which determined the availability of texts for retrieval. The twelve magazines were selected considering their context of production, their communicative purpose, their target audience, and the writers of the articles. As stated by the magazines, they were published by the universities' extension and outreach programs, which aimed at conducting research to develop and offer solutions in response to local needs. As for the communicative purpose, the magazines sought to disseminate research-based information by communicating research results and by offering recommendations and educational resources. The target readers were university students, specialists, and nonspecialists with some expertise in agriculture. The texts were written mainly by faculty and extension specialists.

The period represented in the corpus was restricted to the years 2008 and 2009. As for the use of complete texts and size criteria, a total of 700 full texts of different lengths were assembled, which rendered 455,366 tokens. Another criterion used to ensure representativeness was variety of writers, with a total of 292 authors represented in the corpus. Information about corpus design is synthesized in Table 1.

3.2. Materials, data processing, and analysis

Four applications of *WordSmith Tools* 4.0 (Scott, 2004) were used for the analysis: WordList, Match List, the Auto-join tool, and Concord.

3.2.1. General lexical description of the corpus

In the first stage of analysis, the WordList tool yielded information about corpus size and vocabulary size, computed as number of tokens and types respectively, standardized type/token ratio, and range. As the corpus consisted of 700 texts of different lengths, ranging from 207 to 1,987 tokens, it was necessary to calculate the standardized type/token ratio (STTR), rather than the raw type/token ratio (TTR) in order to correct for differences in text length. The STTR was used in this study for the reasons discussed in the following paragraphs.

Different indices, other than the STTR, have been proposed in order to avoid dependency on text length when measuring lexical variation, such as Yule's *K*, the Zipf slope, U (Uber), Carroll's Corrected TTR (CTTR), LogTTR, and Guiraud's Root TTR (RTTR) (Baayen, 2008; Malvern, Richards, Chipere, & Durán, 2004). These different ways in which the TTR has been transformed to compensate for the token size effect intend to represent mathematically how the TTR falls with increasing token count and intend to turn the TTR into a constant (Malvern et al., 2004). However, despite the efforts to solve the problem of text-dependency, the proposed mathematical transformations do not seem to overcome the sample size effect, as the calculations are not truly independent of the number of tokens and are actually affected by sample size (Baayen, 2008; Malvern et al., 2004).

The *D* measure, proposed by Malvern et al. (2004), is another standardization index used to compensate for differences in text length. The measure is calculated by computing averages of TTRs of random samples. However, random sampling provides slightly different values each time the index is computed. This can be overcome by fixing the number of tokens and calculating the average of sequential sub-samples of the same token size (Malvern et al., 2004). That is precisely the procedure underlying the STTR (Baker, Hardie, & McEnery, 2006; Scott, 2004), the statistical index used in this study.

WordSmith Tools, the software used in this study, computes the STTR as a percentage showing the number of new types for every pre-determined number of tokens across texts of different lengths. In this study, the software was programmed to chunk the corpus and compute the type/token ratio every 200-word segment, which is the minimum text size in the corpus, as suggested by Malvern et al. (2004). As a result, the software calculated the type/token ratio at equally spaced measurement points and then worked out the average of all these measures, thus yielding the STTR index.

The Match List tool was used to classify the words using the AWL (Coxhead, 2000), a version of the GSL (West, 1953) without grammar words, and using a separate list of grammar words, herein called GrWL (retrieved from www.victoria.ac.nz/lals/about/staff/paul-nation). These lists were used so that the results could be comparable to those obtained in previous studies from the analysis of research articles, which have largely used the AWL and the GSL. The GrWL contains 308 function

Table 1
Corpus description.

	Corpus
Genre	Semi-popularization articles
Discipline	Agriculture
Domain	Corn production
Sources of the articles	12 popularization magazines and newsletters published online by 11 American universities
Period (publication date)	2008–2009
Size	455,366 words – 700 texts
Writers	Mostly faculty – 292 writers
Readers	University students, specialists, and nonspecialists with some expertise in agriculture
Purposes	To inform, instruct, explain

words representing determiners (e.g., *a, some, the, this*), auxiliary verbs (e.g., *can, will, should*), prepositions (e.g., *by, of, under, with*), pronouns (e.g., *themselves, yours, we, everything*), coordinators (e.g., *and, or*), subordinators (e.g., *although, if, while*), and a set of 63 cardinal and ordinal numbers in alphabetical form, ranging from zero to nineteen, and then from twenty to one billion in increments of ten. A fourth list was built with the remaining words, called 'Other Words'. After the grouping of words, the corpus coverage of each list was calculated.

3.2.2. High-frequency words

In the second stage of analysis, the high-frequency words were identified and analyzed. To this end, numbers in digital form and grammar words were removed from the corpus wordlist. Grammar words were excluded as they represent a fairly stable number of members which supply grammatical information (Quirk, Greenbaum, Leech, & Svartvik, 1985) and are very likely to occur in different texts, as opposed to lexical words which are closely related to the domain of the texts (Biber, Johansson, Leech, Conrad, & Finegan, 2000).

Following Nation (2005), who states that the dividing line between high-frequency and low-frequency words is based on the researcher's arbitrary decision, I determined the cut-off point by calculating the median, rather than the mean, which was used in other studies (e.g., Hyland & Tse, 2007). I chose the median as this statistic has proved to be a good indicator of central tendency for summarizing data with extreme values (Levine & Stephan, 2010), as in the case of the skewed distribution represented by words of high-frequency at the beginning of wordlists and hapaxes at the end. The median was calculated by dividing the total number of tokens by two to get a middle value so that half of the tokens remained above the median and half below it. I then selected the types whose cumulative frequency covered the tokens above the median. The center of the distribution was signaled by the type ranked in the 354th position, which had a frequency of 144 tokens in the corpus. Thus, the high-frequency words were those whose frequency was equal to or above 144.

Using the Match List tool, the GSL, and the AWL, the high-frequency words were initially grouped into general words, academic words, and 'Other Words', which contained the remaining words. Inflected forms in each list were lemmatized automatically using the Auto-join tool; then derived forms were added manually to the lemmas in order to build word families, as defined by Nation (2001a, 2001b). As pointed out by Sinclair (1991), the headword of each family was the most frequent member. Using the Concord tool, the headwords were observed in context to analyze word use and meaning. The data obtained was used for subsequent semantic categorization of words. In the case of multi-word families, only those derived forms denoting a different meaning from that of the headword, as revealed by the concordances, were analyzed qualitatively for their categorization. Inflected forms signaling number, tense, person, and comparison were not considered for analysis, as they did not represent changes in meaning. To illustrate, in the word family '*seed*' (*seed*, *seedling*, *seeding*, *seeds*), the headword *seed*, and the members *seedling* and *seeding* were analyzed qualitatively, whereas *seeds* was left unanalyzed, as it only represented the plural form of the headword.

Using semantic and pragmatic criteria, rather than quantitative criteria and wordlists available in the literature, the headwords and some derivations were reclassified into technical words (Cabré, 1999) and non-technical words (P. Meyer, 1997). In order to show how technical words acquired meanings as members of a conceptual system specific to the discipline and how non-technical words expressed meanings related to the process of scientific research and communication, both types of vocabulary were sub-categorized based on the categories proposed by Cabré (1999) and P. Meyer (1997). In both cases, ad hoc categories were also created, which emerged from the data.

4. Results

4.1. General lexical description of the corpus

4.1.1. Corpus size, vocabulary size, and standardized type/token ratio

As computed by *WordSmith Tools*, the corpus contained 455,366 tokens and 12,246 types. The software's STTR calculation revealed a 57.71% ratio between types and tokens, indicating that an average number of 57 new types is introduced every 100 tokens in the corpus. This mean frequency can be said to reveal a high lexical variation in the corpus.

4.1.2. Range

Range was observed in terms of word distribution across the corpus texts. Of the total 12,246 types, only 65 appeared in half or more than half of the texts. Of these types, five appeared in all the texts, all of them being grammar words (*and*, *in*, *of*, *the*, *to*). There were 32 items with a range of 475 or more texts, which included mostly grammar words, numbers in digital form, and just one lexical word. By contrast, the vast majority of the words, 11,670 types, appeared in 70 or fewer texts, representing 10 percent or less of the total number of texts in the corpus. Table 2 shows the range of grammar words and the rest of the words across the texts. The table reveals that, as range decreases, the percentage of grammar words decreases, whereas the percentage of the remaining words increases. These results reveal that the words with wide range were mostly grammar words, while most of the remaining words had narrow range.

4.1.3. Coverage of wordlists

There were 3,777 types from the GSL, 1,452 types from the AWL, 249 types from the GrWL, and 6,768 types in the list of Other Words, which included proper names, abbreviations, acronyms, nomenclatures, and numbers in digital form. The

Table 2
Word range in the corpus.

Number of texts	Grammar words (types)	Remaining words (types)	Total number of types
700	5	–	5
	100%		
699–475	30	2	32
	93.75%	**6.25 %**	
474–350	14	14	28
	50%	**50%**	
349–200	34	72	106
	32%	**68%**	
199–100	36	181	217
	16.60%	**83.40%**	
99–71	6	182	188
	3.20%	**96.80%**	
70–35	20	503	523
	3.82%	**96.18%**	
<34	104	11,043	11,147
	0.93%	**99.07%**	

Bold type is used to differentiate graphically percentages from raw numbers.

results indicate that more than half of the types fell into the list of Other Words, outnumbering the GSL, the AWL, and the GrWL combined. As expected, grammar words had the smallest number of types in the corpus, 249, but they provided the greatest coverage, almost 42% of the total running words. The 3,777 content words from the GSL accounted for about 35% of the corpus. Thus, the GSL and the GrWL together provided a corpus coverage of 77%. The 1,452 types from the AWL covered 6% of the total running words. Taken together, grammar words, general words, and academic words accounted for 83% of the corpus tokens. The remaining 6,768 types (Other Words) covered only 17% of the corpus. Table 3 synthesizes the types, tokens, and coverage represented by each list.

4.2. High-frequency words

As revealed by the median, the point in the wordlist separating high-frequency from low-frequency words was the type ranked in the 354th position, with a frequency of 144 tokens, indicating that words had to appear at least 144 times in the corpus to be considered high-frequency items. These results reveal that the high-frequency words, that is the first 354 types in the wordlist, covered 50% of the tokens (above the median), whereas the rest of the types, 11,643 items, covered the remaining 50% of the tokens (below the median). Among the high-frequency words, there were 243 types from the GSL, 29 types from the AWL, and 82 types from the list of Other Words. As Table 4 shows, general words predominated among the high-frequency words of the corpus, whereas academic words were not highly represented among the most frequent words.

4.2.1. Word families of high-frequency words

Following Nation (2001a, 2001b), the 354 high-frequency words were grouped into 277 word families. As indicated in Table 5, most of the word families were single-member families. The majority of multi-member word families had two members. The complete list of word families is presented in Appendix B.

A close look at the word families revealed that many of the high-frequency words in the corpus were directly linked to agriculture, regardless of their initial classification as general, academic, or other words. For instance, general words such as *plant*, *field*, *soil,* and *corn* are items which clearly signal concepts related to the discipline. Likewise, some academic words also denote specialized meanings in agriculture, as in the following cases: *emerge* (e.g., seedling blight occurs after seedlings *emerge;* young nodal roots that *emerge* from the crown area of the plant), *injury* (e.g., typical stresses that can stunt initial nodal development include fertilizer salt *injury*, seedling diseases, herbicide *injury*, insect feeding damage; in field trials, leaf tissue *injury* ratings were similar), *maturity* (e.g., delayed plantings may result in reduced yield and delayed *maturity*; uneven field *maturity* makes swathing a desirable option), and *volunteer* (e.g., timely control of *volunteer* wheat is essential in reducing the spread of wheat streak mosaic disease; the best way to manage *volunteer* corn infestations is to avoid them). These results demonstrate that the initial classification of high-frequency words using the GSL and the AWL did not represent accurately the meanings and uses that the words acquired in agriculture semi-popularization articles.

Table 3
Wordlist types, tokens and coverage in the corpus.

	Types	Tokens	Coverage
GrWL	249	190,857	41.91%
GSL	3,777	158,514	34.81%
AWL	1,452	27,332	6.00%
Other words	6,768	78,663	17.28%
Total	**12,246**	**455,366**	**100%**

Bold type indicates totals.

Table 4
Wordlist types and tokens of high frequency words in the corpus.

	Types	Tokens
GSL	243	96,145
AWL	29	7,671
Other words	82	22,534
Total	**354**	**126,350**

Bold type indicates totals.

4.2.2. Use and meaning of high-frequency words

High-frequency words were re-classified into technical and non-technical using semantic and pragmatic criteria, as proposed by Cabré (1999) and P. Meyer (1997). In order to show the actual meanings and uses of technical and non-technical words in agriculture semi-popularization articles, both types of vocabulary were further sub-categorized using Cabré's (1999) and P. Meyer's (1997) categories as well as categories that emerged from the data.

4.2.2.1. Technical words. Based on Cabré (1999), technical words were defined as terms denoting field-specific concepts semantically related as members of the domain of agriculture. In order to show how technical meaning was determined by the discipline-specific concepts represented by words, the technical words were sub-classified into nine groups that reflect concepts of the discipline: (1) Objects and entities; (2) People; (3) Natural phenomena; (4) Processes, operations, and actions; (5) Properties, states, and qualities; (6) Time; (7) Regions and areas; (8) Chemical nomenclatures; and (9) Units of measurement (Appendix C). Categories 1, 4, and 5 were taken from Cabré (1999). The rest of the categories were created ad hoc based on patterns that emerged from the data. This categorization scheme reveals the actual conceptual meanings of the words as used in one specific discipline, agriculture, and one specific genre, semi-popularization articles. The scheme also provides evidence that technical meaning is determined semantically by the concepts represented by words and pragmatically by the context in which words are used.

(1) The group 'Objects and entities' represented words referring to natural things, material artifacts, and abstract entities, such as plants (1), parts of the plant (2), pests (3), and control instruments (4):

(1) ***Sorghum*** *is notoriously non-responsive to changes…*

(2) *The* ***root*** *cortex is light-gray…*

(3) *…soybean* ***aphids*** *are present in young soybean…*

(4) *…recommended* ***fertilizer*** *application rates…*

(2) The group 'People' included words that referred to individuals involved in activities specific to agriculture:

(5) *Many* ***producers*** *are using earlier planting dates…*

(6) *Some Iowa* ***growers*** *will face special harvest-time issues this fall…*

(3) The group 'Natural phenomena' represented words related to the weather (7) and natural resources (8):

(7) *Probability of* ***frost*** *before maturity…*

(8) *Cultivated soils appear to absorb* ***water*** *readily…*

(4) The group 'Processes, operations, and actions' contained verbs and nominalizations referring to natural events or phenomena (9, 10) and actions performed by people (11, 12):

(9) *Winter annuals that are* ***flowering*** *may require higher rates…*

(10) *This fungus does not cause* ***damage*** *to wheat…*

(11) *Producers may have to* ***harvest*** *wetter corn first…*

(12) *…early* ***application*** *of glyphosate…*

Table 5
High-frequency word families in the corpus.

	GSL	AWL	Other words	Total
1-member families	132	23	59	**214**
2-member families	41	3	10	**54**
3-member families	4	–	1	**5**
4-member families	3	–	–	**3**
5-member families	1	–	–	**1**
Total	**181**	**26**	**70**	**277**

Bold type indicates totals.

(5) The group 'Properties, states and qualities' included adjectives and nouns with a specialized meaning used to describe elements of the discipline, such as entities (13, 14) and actions (15):

(13) …*management of other winter* ***annual*** *weeds*…

(14) *The* ***symptoms*** *in dry bean are similar to soybean*…

(15) …*the impact of* ***delayed*** *planting*…

(6) The group 'Time' contained words referring to moments in which typical agricultural activities are carried out (16) or certain phenomena take place (17):

(16) …*it is important to identify ear rot problems before* ***harvest***…

(17) …*high populations of SCN at the end of the* ***season***…

(7) The group 'Regions and areas' consisted of words representing places where agricultural and research activities are generally performed (18) or where certain phenomena occur (19):

(18) …*collect soil samples from your wheat stubble* ***fields***…

(19) …*Plants in these border* ***areas*** *may be in bloom*…

Following Nation (2001a), proper names were included in Group 7, as their meaning was found to be closely related to the domain of the texts. Names of states such as *Kansas* and *Indiana* represented places where the agricultural and research activities reported in the texts were carried out or where certain phenomena occurred:

(20) …*eight soybean fields across northern* ***Indiana*** *were sampled and tested*…

(21) *Most canola in* ***Kansas*** *is approaching the optimum growth stage*…

(8) The group 'Chemical nomenclatures' included chemical symbols such as *P* (phosphorous), *K* (Potassium), *N* (Nitrogen), and *S* (sulfur). These symbols are relevant in agriculture as they represent soil and plant nutrients:

(22) *Samples for a* ***P*** *and* ***K*** *soil test should be taken to a 6-inch depth*…

(23) *To determine whether* ***P, K, S,*** *and lime are needed on tall fescue and smooth bromegrass pastures or hayfield*…

(9) Finally, the group 'Units of measurement' consisted of three words that represent measures: *inches*, *bu* and *lbs*, used in agriculture to refer to planting practices (24), size of plants (25), and weight/volume (26):

(24) *Seeding deeper than 2* ***inches*** *should be avoided*…

(25) *Once the alfalfa is about 4–6* ***inches*** *tall*…

(26) …*corn with test weights below 50* ***lbs/bu*** *often increases*…

4.2.2.2. Non-technical words. Based on P. Meyer (1997), non-technical words were defined as those used to describe relations among concepts of the discipline, and words related to the research process. Therefore, as opposed to technical words, non-technical words do not express meanings that are strictly discipline-related but refer to meanings that are common to different disciplines, such as those related to the research process and the process of research communication. In order to show these meanings and uses, the non-technical words were sub-classified into six groups, based on P. Meyer: (1) Tense, aspect, and modality; (2) Elements of the subject matter described; (3) Relations between entities; (4) Elements of scholarly practice; (5) The text domain; and (6) Others. Each group was further divided into sub-categories, as proposed by P. Meyer. Some sub-categories were slightly modified and the following were created ad hoc based on patterns that emerged from the data: Time, Properties of entities, Text reference, and Recommendations (Appendix C).

(1) The group 'Tense, aspect, and modality' included words denoting temporal deixis, notions of time, and evaluations about the propositional content of the texts.

(a) Temporal deixis was signaled by the adverb *now* (27) and by some adjectives (28):

(27) *Sudden death syndrome (SDS) is* ***now*** *appearing from the very southern part*…

(28) *With the* ***recent*** *cooler night temperatures, we are seeing an increase in*…

(b) Time reference was represented mostly by adverbs of frequency (29) and verbs which situated events in a point of time (30):

(29) *Potassium deficiency* ***typically*** *appears as a yellowing of leaf edges*…

(30) *Infection can* ***begin*** *before tassel emergence*…

A special group of words that denoted notions of time were nouns which functioned as the head of noun phrases and collocated mostly with technical words to form units denoting field-specific concepts. Although these nouns are not technical terms in a strict sense, when combined with technical words they can be said to convey specialized meanings:

(31) *planting* ***date*** *– seeding* ***date***

(32) *flowering* ***time*** *– harvest* ***time***

(c) Words denoting evaluation were mostly adjectives and adverbs used by the writers to position themselves in relation to the content of the texts:

(33) *This is* ***especially critical*** *on very young corn…*

(2) The group 'Elements of the subject matter described' represented items used to describe entities of the discipline in terms of quantity and other properties. This group included adjectives (34) and nouns (35) that tended to co-occur mostly with technical words:

(34) ***additional*** *nitrogen –* ***high*** *humidity*

(35) ***amount*** *of fertilizer –* ***amount*** *of irrigation*

This group contained a set of nouns referred to as "classifiers of entities" (P. Meyer, 1997, p. 12) representing general types of entities and states of affairs, such as 'activity', 'conditions', 'factors' and 'system'. These nouns resemble the nouns signaling time reference in the group 'Tense, aspect, and modality' in that they regularly co-occured with technical words. They functioned as the head of noun phrases and became specific through the pre- or post-modification of technical words.

(36) *herbicide* ***activity*** *in these weeds*

(37) ***conditions*** *of extreme N-deficiency*

(3) The group 'Relations between entities' included words signaling relationships among specialized concepts, both in the context of the text and in the real world. The words denoting relations among entities in the context of the discourse were used to link different ideas or sections of the text. This group included subordinators such as '*whether*', '*even* if', 'as *long* as', sentence connectors such as '*also*', '*so*', 'in *addition*', and other linking expressions, such as '*due* to' and 'one important *point* to note'. The examples below illustrate some of these cases:

(38) ***Even*** **though** *we have not yet seen any soybean aphids in KY, it is still very early in the season…*

(39) ***As a result****, late planted corn is a riskier crop than soybean…*

The words used to signal relations among entities in the real world included verbs which described cause–effect relations (40), relations of inclusion (41), and other types of relations (42):

(40) *…scab fungus can also* ***cause*** *seedling blight…*

(41) *Symptoms* ***include*** *excess salivation, difficult breathing…*

(42) *…both forage and pods* ***provide*** *digestible protein…*

(4) The group 'Elements of scholarly practice' included items related to the research process. Some of these words described research methods and procedures (43, 44), whereas others referred to different entities (45, 46):

(43) *…the most* ***effective*** *treatments include combinations of dicamba with…*

(44) Table 2 *shows the results of a* ***similar*** *field trial for corn…*

(45) ***Research*** *conducted in neighboring states has also observed…*

(46) *these* ***data*** *are preliminary and further research must be done…*

(5) The group 'The text domain' represented items used to draw the reader's attention to parts of the text, words used to directly address the readers for recommendations and instructions, and words used to perform linguistic acts.

(a) The words used to draw the reader's attention to parts of the text pointed to elements such as figures, graphics, and text sections:

(47) *See the* ***following*** *summaries and comments from these states…*

(48) *The* ***table*** *above presents some University of Nebraska data…*

(b) The second set of words included verbs used to give advice and instructions to the readers:

(49) *…you should* ***consider*** *a treatment in alfalfa less than 10 inches tall…*

(50) ***Use*** *the best quality grain to meet your seed need…*

(c) The third set of words consisted of verbs used by the writers to perform linguistic acts:

(51) *We would* ***point*** *out that locations to our north…*

(52) *Kernel processing […] is strongly* ***recommended*** *for mature corn silage…*

(6) The group 'Others' contained words whose meaning was essentially general, such as *make*, *state*, *extension,* and *university*. These items did not represent concepts specific to agriculture or concepts linked to the meanings and uses of the non-technical words in the corpus, as defined by the categorization scheme used in this study.

5. Discussion

5.1. General lexical description of the corpus

Data about corpus size and vocabulary, range and coverage provided insights into the vocabulary of agriculture semi-popularization articles in English. The variability of distinct words, as revealed by the standardized type/token ratio, and the narrow range of words may be attributed to the large number of texts in the corpus (700) and the length of the texts (an average number of 650 tokens). As observed by Sinclair (1991) and Sutarsyah, Nation, and Kennedy (1994), a variety of short texts allows for more variation in vocabulary than a few long texts.

As for the presence of the AWL in the corpus, the results indicated that the list had a representation in semi-popularization articles different from that of research articles. The 6% coverage of academic words observed in this study for semi-popularization articles is much lower than that reported for research articles: 9.96% (Valipouri & Nassaji, 2013) in chemistry; 10% (Chen & Ge, 2007) and 12.24% (Wang et al., 2008) in medicine; and 11.17% (Vongpumivitch et al., 2009) and 11.96% (Khani & Tazik, 2013) in applied linguistics. It is interesting to note that the 6% coverage of academic words in semi-popularization articles lies in between the 9–12% observed in research articles and the 3.9% reported for newspapers (Coxhead & Nation, 2001). This finding clearly reveals a key feature of the semi-popularization article: the fact that it represents an intermediate genre lying between the research article circulated in academia and the popularization article circulated in the mass media. As for the presence of the GSL in the corpus, the results revealed that the GSL and the GrWL together covered 77% of the corpus tokens, which is slightly higher than the 76.4% coverage reported by Khani and Tazik (2013) for the GSL in applied linguistics research articles, and higher than the 65.46% reported by Valipouri and Nassaji (2013) in chemistry research articles.

An interesting comparison can be made between the presence of items from the GSL and the AWL in the agriculture semi-popularization articles analyzed in this study and in the agriculture research articles analyzed by Martínez et al. (2009). The 6% coverage of academic words observed in semi-popularization articles in this study is lower than the 9.06% observed by Martínez et al. in research articles. In contrast, the 77% coverage of general words in semi-popularization articles is larger than the 67.53% observed in research articles by Martínez et al.

5.2. High-frequency words

The results revealed that many of the high-frequency words signaled specialized meanings, irrespective of their initial classification as general words and academic words. Therefore, in line with previous studies that have questioned the use of the GSL and the AWL for the lexical description of field-specific genres (Hyland & Tse, 2007; Martínez et al., 2009; Neufeld, Hancioglub, & Eldridge, 2011), the findings in this study suggest that the GSL and the AWL also present certain limitations to describe the vocabulary of semi-popularization articles. These limitations seem to be the result of semantic phenomena such as polysemy and homonymy as well as pragmatic phenomena, both of which determine the multiple meanings words may acquire in different contexts. This is particularly relevant in texts that communicate science, such as semi-popularization articles, which include words that acquire meanings revealing the specificity of different disciplines. In agreement with previous studies (Chen & Ge, 2007; Hyland & Tse, 2007, 2009; Khani & Tazik, 2013; Li & Qian, 2010; Martínez et al., 2009; Vongpumivitch et al., 2009; Wang et al., 2008), the present study points to the need to build wordlists from specialized corpora representing specific genres and disciplines.

High-frequency words were reclassified into technical and non-technical using semantic and pragmatic criteria, as suggested by Cabré (1999) and P. Meyer (1997). The approach was very useful in determining the actual meanings and uses of technical and non-technical words, and revealed interesting differences between both types of vocabulary in the corpus. The technical words were found to convey discipline-specific propositional content, and to function as the surface formal expressions of disciplinary concepts. On the other hand, the non-technical words were generally found to (1) evaluate the propositional content – that is, '*best*', '*important*'; (2) guide the reader through the text – as in, 'for *example*', '*following*'; and (3) talk about the subject matter. It is interesting to note that the non-technical words used to talk about the subject matter tended to co-occur in combination with technical words to denote field-specific concepts, such as 'seeding *rate*', 'the *size* of weeds', '*normal* fertilization', and 'insecticide *treatment*'. Although this group of non-technical words did not have a specialized meaning, their meaning can be said to be somehow related to the discipline.

These results provided clear evidence that both semantic and pragmatic features were indicators of technical meaning in the corpus. For instance, general words, such as '*seed*' and '*disease*', and academic words, such as '*emergence*' and '*response*',

were considered technical words because semantically they represented field-specific concepts and because pragmatically they were used in the context of semi-popularization articles. The findings support claims that the specialized meaning of words depends on the semantic relations that the words hold in the conceptual system of a discipline as well as the specialized communicative situations in which the words are used, as pointed out by Cabré (1999) in the field of terminology. It is therefore concluded that frequency criteria may not fully reveal the specialization of the words used in texts that communicate science, such as the semi-popularization articles analyzed in this study.

Finally, the results reveal the usefulness of identifying the vocabulary of scientific genres without using existing wordlists in the literature, as Ward (2009) did for his engineering wordlist, rather than selecting specialized vocabulary on the basis of the GSL and the AWL. Even in the semi-popularization articles analyzed in this study, an intermediate genre between the research articles published in specialized journals and the popularization articles published in the media, general words and academic words acquired technical meanings that clearly reflected concepts of the discipline, in this case agriculture.

5.3. Limitations of the study

Whereas some scholars (Biber, Conrad, & Reppen, 1998) have suggested that more than one million words are needed to make generalizations about word use and meaning, other scholars have observed that small corpora are appropriate for studying high-frequency vocabulary (Hunston, 2002; Kennedy, 1998; Koester, 2010), which was one of the aims in this study. As Koester (2010) points out, in a large corpus, high-frequency items may become unmanageable for analysis, sometimes making it necessary to examine just a sample of the words. However, as claimed by Koester, in a small corpus, all the high-frequency items can be explored, as was done in this study. In addition, as suggested by C. Meyer (2004), small corpora are generally better designed to be approached manually, which was the methodology used for the qualitative analysis of high-frequency words in this study. Therefore, although the sample analyzed represents a small corpus of about 500,000 tokens, it proved to be useful in providing preliminary data about the types of vocabulary that characterize semi-popularization articles in English. Future studies should explore the vocabulary of semi-popularization articles in larger corpora of the same discipline and of other disciplines.

A further limitation is related to the use of word families as a unit of analysis, an issue that has been raised by different scholars (Brezina & Gablasova, 2013; Gardner & Davies, 2013; Ward, 2009). Problems in the use of word families became clear from the semantic and pragmatic analysis of high-frequency words, some of which presented variations in meanings due to polysemy, homonymy, or collocational patterns. Most of the word families in the corpus had only one member, and the few multi-word families identified contained mostly inflected forms. However, in some cases derivations conveyed different meanings from that of the headword. These observations suggest the need to reconsider the notion of word family as a unit of analysis in corpus-based lexical studies, and using the lemma (Nation, 2001a) instead, which contains only inflected forms that do not signal changes in meaning (Brezina & Gablasova, 2013; Gardner & Davies, 2013).

5.4. Pedagogical implications

The study highlights the value of compiling small specialized corpora to build genre and discipline-based wordlists especially tailored to address the needs of learners in specific areas, rather than building universal lists that presuppose fixed divisions of vocabulary common to all disciplines. In particular, the study suggests that wordlist compilation should integrate frequency criteria as well as meaning criteria. Using frequency evidence enables the targeting of the vocabulary that needs to be taught; using meaning evidence enables the capturing of features related to the uses of words, such as the technical meanings that words acquire in specialized contexts (e.g., 'Nebraska research has shown that a *stand* of 90,000 plants/acre will not adversely affect *yield*'), the collocation patterns in which words are typically used (e.g., *harvest aid, harvest height, harvest index, harvest interval, harvest season*), and the fixed multiple-word units that acquire the status of terminological phrases (e.g., *planting date, wheat fusarium head blight, black cutworm, barley yellow dwarf*). The construction of more specialized wordlists, such as the ones compiled in this study, will enable ESP teachers to establish vocabulary goals by addressing not only the question of how many words need to be taught but also how words are used in specific genres and disciplines, such as the agriculture semi-popularization articles analyzed in this study.

The results reported in this study provide insights into the types of vocabulary used in agriculture semi-popularization articles in English. The quantitative and qualitative analysis of the vocabulary enabled the compilation of wordlists based on not only frequency criteria but also meaning criteria. These findings have important pedagogic implications in ESP reading courses lectured at undergraduate level, in which semi-popularization articles are particularly suitable to be exploited pedagogically for teaching reading skills.

6. Concluding remarks

This study represents one step further on the road to a description of the semi-popularization article in English. This is an important contribution to the field, considering that few studies have clearly attempted to analyze the semi-popularization article as a distinct genre. In particular, the present study has shed light on the types of vocabulary used in this genre, an aspect that had not been widely explored before. Future studies should elucidate features of the semi-popularization article at micro and macro levels, as the description of this genre in English still needs considerable development.

Acknowledgments

I am most grateful to Iliana A. Martínez for reading the manuscript. I am also grateful to the anonymous reviewers and this journal's co-editor, Dr Nigel Harwood, for their careful reading of the manuscript and their insightful comments. This study was supported by grants of Consejo Nacional de Investigación Científica y Técnica (CONICET) (No 330/2011, 4749/2013), Argentina.

Appendix A

Table A.1
Magazines, publishing institutions, and number of texts used for the corpus.

Magazine/newsletter	Publishing institution	Number of texts used for the corpus
1 C.O.R.N. Newsletter	*The Ohio State University*	76
2 Integrated Crop Management News	*Iowa State University*	74
3 CropWatch Newsletter	*University of Nebraska*	83
4 Field Crop News	*The Pennsylvania State University – PennState*	85
5 Plant and Soil Sciences Extension Newsletter	*Oklahoma State University*	65
6 Agronomy e-updates	*Kansas State University*	94
7 *Pest&Crop* Newsletter	*Purdue University*	58
8 Crop & Pest Report	*North Dakota State University*	59
9 Weekly Crop Update	*University of Delaware*	28
10 Kentucky Pest News Newsletter	*University of Kentucky*	32
11 Louisiana Agriculture Magazine	*Louisiana State University*	27
12 The Plant Disease Alert Newsletters	*Kansas State University*	19
Total number of texts		**700**

Appendix B

Table B.1
Complete list of word families.

GSL		AWL	Other Words
1. **plant** plants planting planted	13. **grain**	1. **areas** area	1. **soybean** soybeans
2. **fields** field	14. **disease** diseases	2. **emergence** emerge	2. **herbicide** herbicides
3. **soil** soils	15. **year** years	3. **potential**	3. **N**
4. **corn**	16. **weeds** weed	4. **occur** occurs	4. **fungicide** fungicides fungus
5. **application** applied applications apply	17. **high** higher	5. **percent**	5. **moisture**
6. **crop** crops	18. **control**	6. **available**	6. **acre**
7. **wheat**	19. **growth** growing	7. **research**	7. **S**
8. **seed** seedling seeding seeds	20. **early** earlier	8. **injury**	8. **aphids** aphid
9. **yield** yields	21. **late** later	9. **significant**	9. **nitrogen**
10. **use** used using	22. **also**	10. **normal**	10. **feeding** feed
11. **leaf** leaves	23. **low** lower	11. **maturity**	11. **residue** residual
12. **producers** production products product produce	24. **treatment** treatments treated	12. **period**	12. **glyphosate**
25. **season**	42. **cause** caused causes	13. **prior**	13. **insecticide** insecticides
26. **time**	43. **dry** drying	14. **annual**	14. **kernels** kernel
27. **damage** damaged	44. **problem** problems	15. **range**	15. **hybrids** hybrid
28. **increase** increased increases	45. **winter**	16. **factors**	16. **alfalfa**
29. **levels** level	46. **good**	17. **data**	17. **fertilizer**
30. **stage** stages	47. **week** weeks	18. **economic**	18. **nutrient** nutrients
31. **conditions**	48. **result** results	19. **response**	19. **symptoms**
32. **rate** rates	49. **need** needed	20. **similar**	20. **infected** infection
33. **reduce** reduced	50. **development** develop	21. **stress**	21. **Kansas**
34. **days** day	51. **so**	22. **adequate**	22. **scab**
35. **fall**	52. **only**	23. **impact**	23. **species**
36. **varieties** variety	53. **populations** population	24. **volunteer**	24. **sorghum**
37. **harvest**	54. **management**	25. **affected**	25. **foliar**
38. **temperatures** temperature	55. **test**	26. **label**	26. **growers**
39. **loss** losses	56. **ear**		27. **stalk**

Table B.1 (*continued*)

GSL	AWL	Other Words
40. **root** roots	57. **best** better	28. **resistant**
41. **spring**	58. **water**	29. **larvae**
59. **inches** inch	80. **number**	30. **Iowa**
60. **state**	81. **present**	31. **susceptible**
61. **insect** insects	82. **even**	32. **tillage**
62. **stand** stands	83. **effective**	33. **forage**
63. **risk**	84. **especially**	34. **K**
64. **weather**	85. **head**	35. **flowering**
65. **make** made	86. **resistance**	36. **spray**
66. **row** rows	87. **found**	37. **storage**
67. **often**	88. **common**	38. **T**
68. **very**	89. **due**	39. **D**
69. **well**	90. **surface**	40. **F**
70. **small**	91. **rot**	41. **PH**
71. **include** including	92. **cover**	42. **pest**
72. **sample** samples	93. **following**	43. **beetles**
73. **system** systems	94. **reports** reported	44. **P**
74. **likely**	95. **quality**	45. **Nebraska**
75. **see**	96. **cool**	46. **mold**
76. **important**	97. **average**	47. **Indiana**
77. **wet**	98. **then**	48. **BU**
78. **additional** addition	99. **possible**	49. **nitrate**
79. **usually**	100. **long**	50. **frost**
101. **information**	128. **spot**	51. **rainfall**
102. **summer**	129. **recommended**	52. **hail**
103. **rust**	130. **June**	53. **blight**
104. **stem**	131. **just**	54. **scouting**
105. **green**	132. **take**	55. **canopy**
106. **help**	133. **check**	56. **organic**
107. **new**	134. **different**	57. **germination**
108. **provide**	135. **least**	58. **fusarium**
109. **central**	136. **point**	59. **deficiency**
110. **greater**	137. **white**	60. **irrigation**
111. **figure**	138. **extension**	61. **LBS**
112. **generally**	139. **black**	62. **threshold**
113. **date**	140. **delayed**	63. **canola**
114. **determine**	141. **past**	64. **silage**
115. **consider**	142. **compared**	65. **sunflower**
116. **same**	143. **little**	66. **Ohio**
117. **based**	144. **university**	67. **E**
118. **amount**	145. **become**	68. **SCN**
119. **table**	146. **difficult**	69. **drift**
120. **now**	147. **ground**	70. **non**
121. **severe**	148. **keep**	
122. **don**	149. **grass**	
123. **activity**	150. **cost**	
124. **too**	151. **sure**	
125. **mid**	152. **yellow**	
126. **large**	153. **begin**	
127. **weight**	154. **typically**	

Appendix C

Table C.1
High-frequency technical words of the corpus.

Category/group	Words
(1) Objects and entities	alfalfa, aphids, beetles, blight, canola, canopy, *corn*, *cover*, *crop*, *don*, *ear* […mold / …rot], feed (n), fertilizer, forage, fungicide, fungus, fusarium, glyphosate, *grain*, *grass*, *head* [Fusarium…blight / …scab], herbicide, hybrids, *insect*, insecticide, kernels, larvae, *leaf*, matter*, mold, nitrate, nitrogen, nutrient, pest, *plant*, *point** [growing… / black…], *populations*, *products*, residue, *root*, *rot*, *seed*, *seedling*, sorghum, soybean, species, *spot*, *stand* (n), stalk, *stem*, sunflower, *varieties*, *wheat*, *weeds*, *yield*
(2) People	growers, *producers*
(3) Natural phenomena	*air*, frost, hail, *rain*, rainfall, *temperatures*, *water*, *weather*
(4) Processes, operations, and actions	*addition**, *application*, *control*, *damage*, *development*, drift, emergence, feeding, flowering (v)*, germination, *growth*, growing (v)*, *harvest* (v)*, infection, irrigation, *management*, *planting**, *production*, response, scouting, silage, spray, storage, tillage, *treatment**
(5) Properties, states, and qualities	annual, *applied**, *black*, *brown*, *cool*, *damaged*, *delayed*, *disease*, *dry*, *drying*, *early*, *even** (a), flowering (a)*, foliar, *green*, *growing* (a)*, infected, injury, *late*, maturity, *mid*, moisture, organic, *planting**, *planted**, *pressure*, residual, *resistance*, resistant, *risk*, *rust*, scab, SCN, *seeding*, stress, susceptible, symptoms, *treated*, volunteer, *warm*, *wet*, *white*, *yellow*
(6) Time	*August*, *fall*, *harvest* (n)*, *june*, *july*, *season*, *spring*, *summer*, *winter*
(7) Regions and areas	acre, areas, *check** [untreated…], *fields*, *ground*, Indiana, Iowa, Kansas, Nebraska, Ohio, *row*, *soil*, *surface*
(8) Chemical nomenclatures	F, K, N, P, PH, S
(9) Units of measurement	bu, inches, lbs.

Note: * = words classified into more than one category; (n) = noun; (a) = adjective; (v) = verb; *italics*: GSL words; underlining: AWL words; no marking: Other Words.

Table C.2
High-frequency non-technical words of the corpus.

Category	Words
(1) Tense, aspect, and modality	**Temporal deixis:** *now*, *past*, *recent*, *so** [..far] **Time:** *days*, *date*, period, *point** [at this-that-some-any…], *time*, *year*, *week* **Temporal / aspectual properties of processes:** *generally*, *just**, *often*, prior […to], *then*, *typically*, *usually* **Modalities:** *best*, *critical*, *especially*, *even** […more/….when/…with], *good*, *important*, *just**, *likely*, *little**, *matter** [no..], *only*, *possible*, *probably*, *problem*, *so**, *sure*, *too*, *very*, *well* **Beginning of actions:** *begin*, occur, *take* [place]
(2) Elements of the subject matter described	**Abstract quantities or quantitative properties of entities:** - *average* (n), *amount*, *levels*, *number*, percent, range, *rate*, *size*, threshold, *weight* - *average* (a), *additional*, *greater*, *high*, *increased*, *large*, *least*, *limited*, *little**, *long**, *low*, *reduced*, *small* **Quantitative changes of entities:** *increase* (v/n), *loss*, *reduce* **Classifiers of entities:** *activity*, *cases*, *conditions*, factors, impact (n), *information*, potential (n)*, *quality*, *stage*, *system* **Properties of entities:** adequate, available, *central*, *common*, deficiency, *different*, economic, *following**, *full*, *new*, normal, potential (a)*, *poor*, prior, *present**, *same*, *severe*, similar*, *top*, significant
(3) Relations between entities	**In the sphere of discourse:** *addition** [in…], *also*, *due* […to], *even** […if/though], *example* [for… / an…of this], *long** [as…as], *point** [an important… / the…of this / one important…to note], *result** [as a…/as a…of/], *so**, *whether* **Real-world relations between entities:** affected, *associated*, *based*, *become*, *cause*, *compared**, *continue*, *determine*, *found*, *help*, impact (v), *include*, *keep**, *need*, *observed**, *provide*, *result** (v) […in], similar*
(4) Elements of scholarly practice	**Evaluations of theories, methods, procedures, or equipment:** *difficult*, *effective* **Scholarly physical and mental activities and accomplishments, or their results, concerning the subject matter:** *applied**, *compared**, *observed**, *reports*, research, *sample** (v), *shown**, *see**, *study*, *test*, *treatment** **Classifiers of data:** data, *results* (n), *sample** (n)
(5) The text domain	**Speech acts:** point* (v) […out], *recommended* **Text reference:** *figure*, *table*, *shown** **Textual deixis:** *following* **Advice and instructions:** *avoid*, *check**, *consider*, *keep** […in mind / …an eye], *make** […sure], *see**, *sure* [make…], *use** (v)
6) Others	*cost*, *extension*, label, *make**, non, *north*, *state*, *university*, *value*

Note: * = words classified into more than one category; (n) = noun; (a) = adjective; (v) = verb; *italics* = GSL words; underlining = AWL words; no marking = Other Words.

References

Adams Smith, D. E. (1987). The process of popularization. Rewriting medical research papers for the layman: Discussion paper. *Journal of the Royal Society of Medicine, 80*(10), 634-636.
Alcíbar, M. (2004). La divulgación mediática de la ciencia y la tecnología como recontextualización discursiva. *Anàlisi, 31*, 43-70.
Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.

Biber, D. (2008). Representativeness in corpus design. In T. Fontenelle (Ed.), *Practical lexicography. A reader* (pp. 63-87). Oxford: Oxford University Press.
Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press.
Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2000). *Longman grammar of spoken and written English*. Essex: Longman.
Bowker, L., & Pearson, J. (2002). *Working with specialized language. A practical guide to using corpora*. London: Routledge.
Brezina, V., & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the new general service List. *Applied Linguistics*, 1-23.
Cabré, T. (1999). *Terminology. Theory, methods and applications*. Amsterdam: John Benjamins.
Calsamiglia, H. (1997). Divulgar: itinerarios discursivos del saber. Una necesidad, un problema, un hecho. *Quark, Ciencia, Medicina, Comunicación y Cultura, 7*, 9-18.
Calsamiglia, H. (2003). Popularization discourse. *Discourse Studies, 5*(2), 139-146.
Chen, Q., & Ge, G. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes, 26*(4), 502-514.
Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialized texts. *Reading in a Foreign Language, 15*(2), 103-116.
Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System, 32*(2), 251-263.
Ciapuscio, G. (1997). Lingüística y divulgación de la ciencia. *Quark: Ciencia, Medicina, Comunicación y Cultura, 7*, 19-28.
Ciapuscio, G. (2003). Formulation and reformulation procedures in verbal interactions between experts and (semi-) laypersons. *Discourse Studies, 5*(2), 207-233.
Ciapuscio, G., & Kuguel, I. (2002). Hacia una tipología del discurso especializado: Aspectos teóricos y aplicados. In J. García Palacios, & M. T. Fuentes (Eds.), *Entre la terminología, el texto y la traducción* (pp. 37-73). Salamanca: Almar.
Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.
Coxhead, A. (2011). The academic word list 10 years on: Research and teaching implications. *TESOL Quarterly, 45*(2), 355-362.
Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing, 16*(3), 129-147.
Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliqueé, XII*(2), 65-78.
Coxhead, A., & Nation, P. (2001). The specialized vocabulary of English for academic purposes. In J. Flowerdew, & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 252-267). Cambridge: Cambridge University Press.
Dudley-Evans, T., & St John, M. (1998). *Developments in English for specific purposes. A multi-disciplinary approach*. Cambridge: Cambridge University Press.
Gallardo, S. (1998). Estrategias y procedimientos de reformulación en textos de Divulgación Científica. *Revista de la Sociedad Argentina de Lingüística, 1*, 67-79.
Gardner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics, 35*(3), 1-24.
Halliday, M. A. K. (1993). On the language of physical science. In M. A. K. Halliday, & J. R Martin (Eds.), *Writing science: Literacy and discursive power* (pp. 59-68). London: The Falmer Press.
Halliday, M. A. K. (1998). Things and relations. Regrammaticising experience as technical knowledge. In J. R. Martin, & R. Veel (Eds.), *Reading science. Critical and functional perspectives on discourses of science* (pp. 185-236). London: Routledge.
Halliday, M. A. K. (2004). *The language of science*. London: Continuum.
Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
Hyland, K. (2010). Constructing proximity: Relating to readers in popular and professional science. *Journal of English for Academic Purposes, 9*(2), 116-127.
Hyland, K., & Tse, P. (2007). Is there an "Academic vocabulary"? *TESOL Quarterly, 41*(2), 235-253.
Hyland, K., & Tse, P. (2009). Academic lexis and disciplinary practice: Corpus evidence for specificity. *International Journal of English Studies, 9*(2), 111-129.
Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.
Khani, R., & Tazik, K. (2013). Towards the development of an academic word list for applied linguistics research articles. *RELC Journal, 44*(2), 209-232.
Koester, A. (2010). Building small specialized corpora. In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 66-79). London: Routledge.
Levine, D., & Stephan, D. (2010). *Even you can learn statistics*. New Jersey: Pearson Education.
Li, Y., & Qian, D. (2010). Profiling the academic word list (AWL) in a financial corpus. *System, 38*(3), 402-411.
Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Hampshire: Palgrave Macmillan.
Mapelli, G. (2004). Estrategias lingüístico-discursivas de la divulgación científica. In *Escritura y conflicto: AISPI Actas XXII* (Vol. II, pp. 169-184)Centro Virtual Cervantes. Retrieved August 19, 2009, from http://cvc.cervantes.es/literatura/aispi/pdf/20/II_12.pdf
Martínez, I., Beck, S., & Panza, C. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes, 28*(3), 183-198.
Mayor Serrano, B. (2003). Elementos metacomunicativos en el artículo de divulgación médica (inglés-español) e implicaciones didácticas para la formación de traductores. *Ibérica, 6*, 89-107.
Meyer, P. G. (1997). *Coming to know: Studies in the lexical semantics and pragmatics of academic English*. Tübingen: Gunter Narr.
Meyer, C. (2004). *English corpus linguistics. An introduction*. Cambridge: Cambridge University Press.
Myers, G. (1990). *Writing biology. Texts in the social construction of scientific knowledge*. London: The University of Wisconsin Press.
Myers, G. (1994). Narratives of science and nature in popularizing molecular genetics. In M. Coulthard (Ed.), *Advances in written text analysis* (pp. 179-190). London: Routledge.
Myers, G. (2003). Discourse studies of scientific popularizations: Questioning the boundaries. *Discourse Studies, 5*(2), 265-279.
Nation, P. (2001a). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
Nation, P. (2001b). Using small corpora to investigate learner needs. In M. Ghadessy, A. Henry, & R. Roseberry (Eds.), *Small corpus studies in ELT. Theory and practice* (pp. 31-45). Amsterdam: John Benjamins.
Nation, P. (2005). Teaching and learning vocabulary. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 581-595). Mahwah: Lawrence Erlbaum.
Neufeld, S., Hancioglub, N., & Eldridge, J. (2011). Beware the range in RANGE, and the academic in the AWL. *System, 39*(4), 533-538.
Pearson, J. (1998). *Terms in context*. Amsterdam: John Benjamins Publishing Company.
Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Essex: Longman.
Schmitt, N., & Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 1-20.
Scott, M. (2004). *WordSmith tools* (Version 4.0) [Software]. Oxford: Oxford University Press.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Sinclair, J. (2005). Corpus and text: Basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1-16). Oxford: Oxbow Books. Retrieved December 9, 2009, from http://ahds.ac.uk/linguistic-corpora/
Sutarsyah, C., Nation, P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus-based study. *RELC Journal, 25*(2), 34-50.
Swales, J. M. (1990). *Genre analysis. English in academic and research settings*. Cambridge: Cambridge University Press.
Valipouri, L., & Nassaji, H. (2013). A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes, 12*(4), 248-263.
Vongpumivitch, V., Huang, J., & Chang, Y. (2009). Frequency analysis of the words in the academic word list (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes, 28*(1), 33-41.
Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academic word list. *English for Specific Purposes, 27*(4), 442-458.

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, *28*(3), 170-182.
West, M. (1953). *A general service list of English words*. London: Longman.

Verónica L. Muñoz holds an M.A. in Applied Linguistics and is a PhD candidate in the same field. She is a teacher assistant in grammar (School of Humanities) and in ESP reading courses (School of Engineering) – UNRC. Her main research interests include ESP, corpus linguistics and genre analysis.