From Frequency to Instructional Order: Insights from a Narrow-angle Corpus of Psychology RA Introductions

Massoud Yaghoubi-Notash
Department of English Language & Literature, Faculty of Persian Literature & Foreign Languages, University of Tabriz, Tabriz, P O BOX 51665331, Iran

Massoumeh Janghi-Golezani
Department of English Language & Literature, Department for Post-graduate Studies, Faculty of Humanities, Islamic Azad University, Tabriz Branch, Tabriz, Iran

Abstract—For various reasons, ESP/EAP is credited with a good record of corpus-based inquiry. Among many different units in the body of language as a quantifiable pool, vocabulary can be an essential category for corpus-based studies. This study cuts across findings that deal with specialized or academic vocabulary, and presents a frequency-based set of lexical words derived from a narrow-angled corpus of psychology RA introductions. For the purpose of the study, 200 research article introductions in psychology were examined for their frequency obtaining 20 high frequency lexical words. Subsequently, the sentential contexts where each high frequency word occurred were put together and examined for difficulty. This provided an empirical basis for the instructional order of vocabulary items regarding the difficulty of the context in which each high frequency lexical word occurred. Findings from the study can cast light on which lexical words and with which instructional order can be instructed to meet the pedagogic needs of learners in ESP/EAP contexts.

Index Terms—corpus, ESP/EAP, frequency, instructional order, psychology RAs

I. INTRODUCTION

Academic writing, as influenced by the sweeping developments in science and technology, is now turning into "a great challenge for non-native speakers of English (NNSs) to participate actively in the international discourse community." (Chang and Kuo, 2011, p. 222)

Such a demand, "in the 'publish-or-perish' academic culture" (ibid, p.222), has stimulated a growing interest in literature to characterize EAP texts as realizations of the academic discourse in their related international discourse community. Down to its barebones, such a discourse is assumed to be initially concerned with scholarly lexical choices. Since an outstanding (if not to say a distinguishing) feature of ESP/EAP "is a heavy load of corresponding specialized vocabulary" (Chujo & Utiyama, 2006, p. 256), vocabulary instruction can take on a central importance as one of primary instructional goals.

On the other hand, in our daily lives, we may encounter a large amount of words or linguistic patterns in various written texts that we have to interact with, especially those which are academically essential for the reasons stated above. Texts, among other things, are quantitative realizations of linguistics choices, and in being so, they are representative of the dominant patterns of use. Therefore, instructional and/or pedagogical approaches to language (and EAP in particular) can be expected to benefit from a time efficient way of exploring some of the essential patterns of texts (as realizations of discourse) in a given field (see Reppen, 2010).

Corpus linguistics is, in effect, the very field to address the issues of time efficient explorations in texts. Corpus linguistics is involved the collection and analysis of large amounts of naturally occurring spoken or written data in electronic format, “selected according to external criteria to represent, as far as possible, a language or language variety as a source of linguistic research” (Sinclair, 2004, p. 16). According to McCarthy and O’Keefe (2003), the real emergence of corpus linguistics started with the revolution of computer in the 1980s and 1990s. In fact, earlier corpora were not computerized. These earlier corpora were called pre-electronic corpora. Kennedy (1998, p.13) noted that “[c]orporus-based research that occurred before the 1960s is generally categorized as pre electronic and consisted of work in five main areas: biblical and literary studies, lexicography, dialect studies, language education studies and grammatical studies”. Such research was often carefully carried out, using index cards and human effort to calculate frequencies by counting.

Although the roots of corpus linguistics can go further back, the real appearance came with the access to machine readable texts which could be stored, transported, and analyzed electronically. Computer corpora are analyzed with the help of software packages such as concordance (Scott, 2004), which includes a number of text-handling tools to support
quantitative and qualitative textual data analysis. As stated by Kenedy (1998), before the 1960s, the analysis of huge bodies of text ‘by hand’ can be burdensome and is not easily replicable. So, computers became an important tool for linguists to be less dependent upon computational expertise. Kennedy adds that corpus linguistics is closely linked to computer by unbelievable speed, total accountability, accurate replicability and the ability to handle huge amount of data. In a similar line of argument, Gavioli (2005) emphasizes the potential relevance of corpus linguistics for language teaching adding that pedagogy attempts to reduce the time that would be necessary to learn a language through a) exposure alone, and b) potential usefulness and likelihood of occurrence.

Corpus-based research as perhaps a unique approach to ESP/EAP text characterization in terms of lexical content can bear implications for instructional purposes (see Flowerdew, 2002). Such a trend of research typifies a) a large body of authentic materials, b) data-driven, probabilistic models, c) automatic or semi-automatic text analysis, and d) contextualized language use (Chang and Kuo, 2011, p. 223). Research and the needs of various sorts have prompted the development of different types of corpora according to (Pearson, 1998):

a) Specialized corpus: Specialized corpora is devised with more specific research goals in mind and focuses on a particular spoken or written variety of language. A kind of specialized corpus that is important for language teachers is learner’s corpus.

b) General corpus/reference: Such as the Brown Corpus, the LOB Corpus, general corpus aims to represent language in its broadest sense and to serve as a available resource for baseline or comparative studies of general linguistic features. General corpora are designed to be quite large. A general corpus is intended to be balanced and include language samples from a wide range of registers or genres.

c) Multilingual Corpus: A corpus that contains texts in more than one language is a multilingual corpus. An example is the Enabling Minority Language Engineering (EMILLE) corpus.

d) Parallel Corpus: A parallel corpus consists of two or more corpora that have been sampled in the same way from different languages. The prototypical parallel corpus consists of the same documents in a number of languages that is a set of texts and their translations. Since official documents (technical manuals, government information leaflets, parliamentary proceedings etc.) are frequently translated, these types of text are often found in parallel corpora.

e) Learner Corpus: A learner corpus consists of language output produced by learners of a language. Most learner corpora consist of written essays using pre-set topics produced in language-teaching classrooms.

f) Diachronic Corpus: A diachronic corpus is a corpus that has been carefully built in order to be representative of a language or language variety over a particular period of time, so that it is possible for researchers to track linguistic changes within it.

g) Dynamic/Monitor Corpus: A dynamic corpus is one which is continually growing over time, as opposed to a static corpus, which does not change in size once it has been built. Dynamic corpora are useful in that they provide the means to monitor language change over time – for this reason they are sometimes referred to as monitor corpora.

A. Corpus and Vocabulary

The levels of information that can be gathered from a corpus range from simple word lists to categories of different complex grammatical structures. Analyses can explore individual lexical and linguistic features or identify clusters of features in and across the texts that characterize particular register (Biber, 1988). Schmitt (2002) states that the basic information to be obtained from a corpus, is the information about the frequency of occurrence. A word list is simply a list of all the words that occur in the corpus, a collection of which can be arranged in alphabetic or frequency order (from most frequent to least frequent). Word lists derived from corpora can be useful for vocabulary instruction and test development.

A majority of corpus work has provided the criteria to generate ‘specialized word lists’ specified to particular genres (Sinclair, 1996). With regard to the large number of vocabulary items in a language, researchers have produced word lists of the recurrent vocabulary in academic texts to maximize the effectiveness of vocabulary instruction. These lists are thought to provide the vocabulary necessary to function in academic contexts (see Coxhead, 2010 for example). According to Chen & Ge (2007), words in academic writing can be divided into four categories:

a) High-frequency words, which those basic English words that constitute the majority of colloquial conversation or speech as well as all the running words in all types of writing. Language learners/users have plenty of chances to get exposed to these words.

b) Technical words, which are the ones used in a specialized field, considerably different from subject to subject.

c) Low-frequency words that are rarely used terms.

d) Academic words, which are somewhere in between the high-frequency words and technical words and have some important functions and account for a relatively high proportion of running words in all academic texts; and acquiring these words seems to be essential when learners are preparing for English for Academic Purposes (EAP).

B. Findings on Corpus-based Works

Corpora have been used in EAP since the 1980s, mainly for research, but a growing number of researchers and practitioners have been advocating the use of corpora in EAP pedagogy. More recently, however, corpus tools and corpus evidence have not only been used as a basis for linguistic research, but also in the teaching and learning of languages. Fundamentally, corpus linguistics has had a strong link with language teaching. John Sinclair’s impact on
First of all, twenty top frequency lexical words were chosen for frequency and determining the instructional order. This number of lexical words as the focus of the study was because most EP teachers who were consulted agreed that
on average twenty words can be taught in one classroom session. Furthermore, practicality concerns especially with determining their context-based difficulty necessitated a maximum of 20 words to be examined.

For the second stage, all two-hundred introductions from their corresponding files were cut and pasted (pdf format) to Microsoft Word 2007 file. The resulting bulk was then fed into the software used in this study (CONC330) to detect and list the words in the ascending order of frequency. Function words were ignored and only twenty (content) lexical words were included in ascending order. Then, each of the twenty high frequency words (HFWs) was located in its corresponding sentence in the bulk. All sentences containing a particular high HFW were copied and pasted into a separate Microsoft Word file. Therefore, there were ultimately 20 files each containing bulks of sentential contexts corresponding to each high-frequency word.

III. Data Analysis

As far as the R.Q.1., i.e. “What vocabulary items are typically used with a higher frequency in psychology research article introductions?” is concerned, the following pattern of high frequency words could be obtained (see Table 4.1).

<table>
<thead>
<tr>
<th>Table 4.1</th>
<th>Top 20 Highest Frequency Lexical Words in Psychology Introductions in Descending Order</th>
</tr>
</thead>
</table>

As Table 4.1 illustrates, the word 'social' is the highest frequency word that occurs 1170 times in local corpus, and the word 'found' is the lowest frequency word occurring 299 times. In between are the words 'have' (945), 'has' (715), 'research' (522), 'behavior' (531), 'children' (523), 'psychology' (515), 'school' (457), 'study' (448), 'other' (444), 'studies' (421), 'students' (406), 'information' (352), 'group' (351), 'positive' (320), 'theory' (317), 'learning' (315), 'different' (310), and 'people' (314).

In order to answer the second research question, that is ‘Which sequential order of instruction can be derived from the psychology introduction corpus?’ each of the twenty words were located in the sentence. Then, all sentences carrying the word in question were put together making a bulk of sentential contexts for each word. So, there were 20 bulks for twenty high frequency words. The readability level of each bulk (as the aggregate of sentential contexts surrounding each word) was calculated. Readability values appear in Table 4.2.

| Table 4.2 | Readability Value of the Aggregates (in Ascending Order) of the 20 High Frequency Words |

As Table 4.2 illustrates, the word 'studies' has the lowest grade-level as indicated by the Flesch-Kincaid Grade Level (15.2). It means that the word 'studies' has the easiest context to read. Conversely, the word ‘psychology’ has the
highest bulk readability level (19) indicating that it occurs in the most difficult context. Between the easiest and the most difficult context-related words these words occur respectively: 'has' (15.6), 'have' (15.8), 'people' (15.9), 'other' (16), 'students' (16), 'research' (16.2), 'learning' (16.5), 'school' (16.6), 'children' (16.8), 'social' (16.9), 'found' (16.9), 'information' (17), 'different' (17), 'study' (17.3), 'behavior' (17.8), 'positive' (18), and 'theory' (18.2).

IV. DISCUSSION & CONCLUSION

Findings on high frequency words provide us with a picture of lexical words that appear most frequently in the introduction sections of psychology RAs. Such a distribution, although apparently a mere count of words, gives us the clue that the seeming general words do have a certain frequency of occurrence in the psychology article introductions. The twenty high frequency words are all general lexical words that may appear in general texts as well. However, they very frequency distribution of each word is revealing enough since it characterizes the unique way in which ideas in this field (i.e., psychology) are communicated through the load of lexical elements. Therefore, the findings on frequency distribution are confirmed in broad terms by Coxhead (2000). Also, from a methodological point of view, this part of the study is well-supported by Hafner & Candlin (2007), and Tribble (2010) who defended the use of localized, field-specific micro-corpus. This study appears to corroborate the limited usefulness of academic word list as argued by Chen & Ge (2007), Hyland & Tse (2007), Martínez et al. (2009) and Paquot, (2007), and Krishnamurthy & Kosem (2007).

In terms of the quantification measures, this finding, although admittedly very simple regarding the mathematical approach, can be one type of studies alongside Kennedy (2003), and Nelson (2000). If one assumes that EAP learners would have to learn (and ultimately use) the word in its context, and the order of difficulty can best suit learners’ ‘easy-to-difficult’ route, then the second finding can be seen as consistent with Krishnamurthy & Kosem (2007) who address the learner-centered aspects of corpus. To the best of author’s knowledge, no similar study has ever addressed the issue of the instructional order.

The second finding of the study offers an empirical framework for determining the instructional order as derived from the bulk of sentential contexts for each high-frequency word. It can be argued that high frequency lexical words are expected to be instructed for the purpose of writing (if not to say for vocabulary learning). In order to so, the instructor (EAP teacher) will be required to follow an order of presentation. Following this finding, such an order can be determined through presenting the word with an easier context first, and then continuing to teach those words that occur in more difficult (sentential) context.

In general, it can be claimed that dividing the words into general, academic, and specialized may not be a completely helpful classification at least when the learner is involved in the writing process. ESP/EAP texts may simply be so not necessarily because of their academic or specialized vocabulary content, but because the so-called general words have a certain (field-specific) pattern of frequency distribution.

Identifying the high-frequency words through corpus-based analysis is a time-efficient practice in line with the nature of ESP and EAP pedagogy. Besides, ordering the words from those in the easiest context to those in the most difficult context provides a psycholinguistically sound basis for instruction since it is believed that the learners would developmentally be ready for the easiest learning prior to tracing the learning path the most difficult. All these have clear implications for ESP/EAP teaching, assessment, and syllabus design where ordering, sequencing, and grading have been generic problems.

REFERENCES

Massoud Yaghoubi-Notash was born in Tabriz, Iran in 1975. He got his B.A. in Teaching English as a Foreign Language (TEFL) from Islamic Azad University-Tabriz Branch in 1998. In 2001, he completed his MA studies in ELT at the University of Tabriz where he graduated with a doctoral degree in ELT in 2007. He is currently an assistant professor of ELT and a full-time member of English Department at the University of Tabriz. He is currently the advisor to the International Relations Office of his affiliated university. Dr. Yaghoubi-Notash has published articles in international journals (e.g. Issues in Applied Linguistics, KOREA Tesol, JLTR, etc) and presented articles in international conferences such as AILA 2008, AIAL 2011, and so forth. His areas of interest are gender and language, task-based language teaching, ESP/EAP, and discourse analysis.

Massoumeh Janghi-Golezani was born in 1981 in Slamas, Western Azerbaijan Province, Iran. She has got an A.Sc in Medical Laboratory Science from Islamic Azad University-Marand Branch in 2003. In 2010, she received her B.A. in Teaching English as a Foreign Language (TEFL) from the Islamic Azad University-Salmas Branch. In 2012, she graduated from Islamic Azad University-Tabriz Branch with a master’s degree in TEFL.