Application of Corpus-Based Text Statistics Analyzer in Sports English

Beibei Yu*
Shandong Sports Institute, Jinan, Shandong, China

*Corresponding author e-mail: beibeiyu@sdpei.edu

Abstract. The link between corpora and ESP is crucial and even more for some time to come, due to the current patterns of academic and professional mobility which create a need for ESP expertise and teaching materials. The problem in Sports English in China since 2004 is the lack of competent teachers and proper-leveled textbooks, which hindered the development of the major of Sports English greatly. This essay discusses four corpus-based online text analyzers to offer statistics of individual texts and vocabulary profile to help English teachers choose proper-leveled texts and modify materials to meet their needs in Sports English teaching.

Keywords: Online Corpus Tool, Text Analyzer, Vocabulary Profile, ESP Textbook Compiling

1. Introduction
Corpus linguistics has exerted a considerable influence in the field of language teaching. The ways in which corpora have been employed in language pedagogy can be divided into two main categories: indirect and direct applications. In the former, corpora are used to inform the design and development of syllabuses, tests and teaching materials, while in the latter corpus data are used for data-driven learning (DDL); that is, hands-on activities in which learners themselves engage in corpus analysis. This essay aims to discuss the applications of text statistics analyzer in developing teaching materials like textbooks in Sports English.

2. Corpus and ESP
Languages for specific purposes (LSP) refers to the area of applied linguistics devoted to the study, description and teaching of different types of language used by specific discourse communities [1]. ESP is defined as “language research and instruction that focuses on the specific communicative needs and practices of particular social groups” [2]. Two main areas that are of interest to ESP researchers are English for academic purposes (EAP) and English for occupational purposes (EOP) such as varieties of English used in medical, agriculture, legal or business contexts. There is no denying that Sports English is a branch of ESP.

A corpus-informed approach to teaching and materials design is seminal in ESP, as it is not unusual for English language teachers to find themselves teaching English for specialisms of which they have no particular knowledge. English teacher are not necessarily interested in and good at sports. And they
are not able to know about every sports event. Sports teachers are expertise in sports, but cannot be competent more because Sports English is anyway a branch of Applied Linguistics which require more knowledge of ESL (English as a Second Language). Students who learn Sports English are majored in sports but have low level (lower than B1) of English proficiency. Therefore, Sports English major has been established in ten sports institutions in China since 2004, it is still gloomy and still an area that clearly deserves our attention from both a research and practitioner perspective.

Three main uses of corpora in ESP are identified by Hewings [3]: First, in conducting research that underpins effective language teaching and learning. Second, corpora can be used as a way of presenting authentic language data to learners in an approach referred to as ‘data-driven learning’ (DDL). Third, corpora can be a resource for teachers to use in preparing classroom teaching activities and materials.

3. Corpus-Informed Teaching Materials

Teaching materials such as textbooks or self-study books are a key source of language input L2 learners are exposed to, in particular in contexts where English is taught as a foreign language. Therefore, it is of utmost importance that the language found in such materials reflects real-life communication that takes place in natural settings rather than be contrived solely for the purpose of covering a rigid teaching syllabus. Corpus-based research should inform the development of textbooks and this is true for both expert user and learner corpora. As Gilmore observes, “textbook authors are not yet habitually checking their materials against relevant corpus data” [4]. This results in the fact that many features that characterize natural discourse are not highlighted in textbooks well enough and the language that learners are exposed to differs considerably from the actual usage. This can be attributed to two main reasons: a lack of awareness of the benefits of corpora among teachers and materials writers, and the amount of work that is necessary to translate corpus findings into accessible teaching materials. For Sports English teachers, it is especially important to avoid pure intuition and stilted examples with limited sports knowledge when choose teaching materials. Meanwhile, difficulty of textbooks could be adjusted [5].

4. Frequency, Lexical Overage and Bands of Vocabulary

One of the most basic applications of corpora is ranking words on the basis of their frequency of occurrence. By analyzing corpora teachers can make fairly reliable statements about how many words are ‘in circulation’ in everyday communication”. What is more, such empirically derived statements about tendencies in language are more accurate than those based on speakers’ intuitions. Specifically, research demonstrates that intuitions of both native and non-native speakers are often unreliable, in particular when it comes to judging lower frequency vocabulary, which is often technical terms in an ESP.

Dividing vocabulary into categories such as high-, mid- and low-frequency is an example of the usefulness of frequency analysis. It is generally agreed that high-frequency English vocabulary consists of 3,000 of the most frequent word families. Since the most frequent words are used to create different texts and they play a key role in successful language use, many scholars treat them as core vocabulary that needs to be acquired as quickly as possible. As explained by Nation [6], high-frequency words are “a relatively small, very useful group of words that are important no matter what use is made of the language”. The significance of high-frequency words is corroborated by research on lexical coverage. Lexical coverage is usually defined as the percentage of words that are known in a given piece of discourse. It is an important notion because it allows us to estimate how many words are needed for successful comprehension of written and spoken texts and making such calculations is greatly facilitated by corpus analyses. As contended by O’Keeffe et al. [7], corpus evidence can be used to assess how many words a reader/listener needs to know to understand a given proportion of any written or spoken text.

Crucially, words differ in terms of how much coverage they provide, with high frequency vocabulary accounting for the largest amount of text. As O’Keeffe et al. observe, “the first 2,000 or so
word-forms do most of the work” and cover “more than 80% of all of the words in spoken and written texts” [7]. In a similar vein, Dang and Webb [8] state that high-frequency vocabulary offers learners “a good return for their learning effort” because this relatively small group of words cover large amounts of different kinds of texts. Words at lower frequency levels, in turn, provide a progressively smaller coverage by dint of their decreasing frequency of occurrence. With regard to this issue, Schmitt and Schmitt [9] point out that beyond the 2–3,000 frequency levels the frequency of occurrence drops off to low levels and this is consistently confirmed by data from a range of different corpora. Thus high-frequency vocabulary can be defined as that which occurs before the coverage percentages become small. This figure is more valid pedagogically and given the importance of such high-frequency vocabulary, learning it should be treated as a milestone in language development. Table 1 shows how much coverage is provided by the different types of vocabulary as attested by data from the BNC. It is worth noting that the figure includes the category of mid-frequency vocabulary, the new construct proposed by Schmitt and Schmitt (2014) as a way of reassessing the traditional frequency-based bands.

Table 1. Coverage Levels for Different Types of Vocabulary (Nation and Anthony 2013)

| Type of vocabulary (word family frequency level) | Coverage (%) |
|------------------------------------------------|--------------|
| High-frequency (1st–3rd 1,000)                  | 90           |
| Mid-frequency (4th–9th 1,000 onwards)          | 5            |
| Low-frequency (10th 1,000 onwards)             | 1-2          |
| Other (e.g. proper nouns or abbreviations)      | 3-4          |
| Total                                           | 100          |

According to Schmitt and Schmitt [7], the introduction of this new category is pedagogically motivated and mid-frequency words deserve the attention of both teachers and learners. The authors estimate that mid-frequency vocabulary consists of around 6,000 word families that lie between the first 3,000 (a threshold for high-frequency vocabulary) and 9,000 (a threshold for low-frequency vocabulary) most frequent words in English. Dang and Webb [9] rightly observe that reaching these levels of lexical competence is rather unlikely for many L2 learners, as they often struggle with learning even the most frequent items. Consequently, the authors suggest that perhaps the first 1,000 words should be treated as a ‘general service’ list and prioritized in L2 programs, as learning this high-frequency vocabulary seems like “a more attainable objective in the EFL/ESL context” [9]. Of course learning these words might still take a long time but knowing them is an essential condition of successful L2 use and can open up new possibilities by more efficient and enjoyable reading or accessing a wider range of texts. Low-frequency vocabulary is perceived as words that are beyond the 9,000 frequency threshold, often regarded as “restricted to certain subject areas” [6]. This 9,000 threshold is taken from Nation’s study in which the author found that L2 learners need a receptive lexical knowledge of 8–9,000 word families to reach a 98% coverage.

As the author calculates, with this level of coverage, forty-nine out of every fifty words in a text should be known, which allows a pleasurable engagement with and understanding of authentic texts. As stated by Schmitt and Schmitt, “if 8–9,000 word families is enough to enable both listening to and reading a wide range of texts . . ., then the low-frequency/utility vocabulary can plausibly be defined as anything beyond this frequency level”[9]. Moreover, low-frequency vocabulary is frequently defined as specialized and technical words which are important for understanding specialized texts in a particular subject area. Such words can be identified by comparing wordlists from general and specialized corpora. That is, technical words tend to be more frequent in a specialized corpus while in large general corpora, which represent language as a whole, such words come up as low-frequency items. Thus, by relying on frequency as a key factor that organizes the classification of vocabulary,
one can identify different types of technical vocabulary like words that are characteristic of specific topics or fields of study.

5. Frequency-Based Analysis of Vocabulary Load

5.1 VocabProfile at Lextutor

Having explained how frequency information can be used to categorize words on the basis of their occurrence, the writer needs to discuss details of how such an analysis can be conducted. The procedure is called lexical frequency profiling (LFP) and Laufer and Nation (1995) were the first authors to use it as a measure of the lexical richness or sophistication of texts. The LFP can be carried out by means of a corpus program called VocabProfile. The program is available on Tom Cobb’s (2015) Lextutor via http://www.lextutor.ca/vp/comp/. Firstly, paste a text that you want to use for teaching purposes into the input box and click “submit”. Then it will produce a lexical profile of your text. It classifies all words from a given text into four frequency levels: the first 1,000 most frequent word families (1K band), the second 1,000 most frequent word families (2K band), academic vocabulary (identified on the basis of Coxhead’s (2000) Academic Word List) and an off-list which contains all the remaining vocabulary. However, the most recent version of the program, BNC-COCA-25, analyzes texts at twenty-five levels of frequency. These levels are based on twenty-five 1,000-word bands created by Paul Nation (http://www.lextutor.ca/vp/comp/coca.html). Except for dividing words into the twenty-five frequency levels, the program also provides other types of frequency information such as tokens, types, cumulative tokens and a type-token ratio.

Fig 1. Example of output from VocabProfile available at Lextutor

According to Cobb and Horst, “a typical profile for written English texts is 70% items from the most frequent 1,000 word families, 10% from the second, and the remainder from less frequent zones”[9]. However, these frequency distributions might vary depending on the type of texts that are chosen. For instance, academic texts are likely to contain a higher degree of lower-frequency vocabulary than informal texts.

Another tool on the same website is RANGE which is used to analyze the distribution of words across different texts or sections of corpora. It is a free program which can be downloaded from Paul Nation’s website (http://www.victoria.ac.nz/lals/about/staff/paul-nation) or accessed online via Lextutor (http://www.lextutor.ca/range/). The program is particularly useful when you wish to compare a series of texts from a textbook or graded reader because it can be used to explore the number of repetitions of individual words throughout the text under study. Such information is valuable because it helps you determine whether the text provides enough repetition of vocabulary for incidental learning to occur. RANGE operates in such a way that it creates the lexical profiles of texts that are uploaded and compares them with reference wordlists derived from large corpora which is the same lists as the ones used in VocabProfile. Through this procedure, you can obtain detailed information about the analyzed texts in terms of the number of tokens, types and word families they contain [10].

5.2 Text Statistics Analyzer
Another tool worth exploring in relation to vocabulary profiling is Text Statistics Analyzer at https://www.usingenglish.com/resources/text-statistics/. The basic function of this online statistics tool Text Analyzer is to get “Detailed Text Statistics” to research the structure of your text. It will show you statistics about your text to help you understand its complexity and readability. Perfect for use by students, translators, writers and anyone wanting to understand their text statistically: Word Count; Unique Words; Number of Paragraphs; Number of Sentences; Words per Sentence; Number of Characters; Characters per Word; Number of Syllables; Syllables per Word.

The second function of Text Analyzer is to test “Readability” to discover how understandable your text is. Use these readability statistics to help you assess the complexity of a text and how hard it is to read and understand. These industry-standard tests are designed to give you a statistical analysis of the difficulty of your text, allowing you to see if it's going to engage with your desired audience. This version of our text analyzer comes with the following tests: Hard Words; Long Words; Lexical Density; Gunning Fog Readability Index.

![Fig 2. Part example of output of Text Analyzer](image)

The advanced function of Text Analyzer is to get even more statistics about your text. It gives a much more detailed analysis of text with many more statistics, analysis by word length, by word frequency and by common phrases. Advanced Features Include: Save up to 20 texts for future reference and comparison; Analyze longer texts of up to 100,000 characters; A Graded Text Analysis Tool; A fully sortable Word Frequency List; A complete breakdown of the text by Word Length.

Meanwhile, a “Word cloud” will be generated, which is a graphical representation of word frequency from a text. The size of a word shows how important it is and how often it appears in a text — its frequency. It could capture and display the overall theme of the book based on nothing other than an ‘eye-ball’ of what’s frequently repeated in the text. People typically use word clouds to easily produce a summary of large documents (reports, speeches), to create art on a topic or to visualize data.

5.3 www.wordandphrase.info

Yet the third tool creates vocabulary profiling is the AnalyzeText option available at http://www.wordandphrase.info/analyzeText.asp. The tool has been created by Mark Davies and it allows you to analyze different texts on the basis of frequency information from COCA.

There will be three percentages in the table above show the number of words in your text in each of three different "frequency ranges", which are based on data from COCA. Words #1-500 in that corpus are RANGE 1, whereas RANGE 3 are low frequency words. The color coding is conducive for one to see quickly and easily which words in your text might be new for you or provide insight into the topic of the text. An ESP instructor can further categorize the subject-specific vocabulary based on his own content knowledge. He could adjust the difficulty of general English and pick up the technical terms at the same time.
5.4 AntWordProfiler

This is also a free piece of corpus processing software that similar analyses can be conducted, which can be downloaded from Laurence Anthony’s website: http://www.laurenceanthony.net/software/antwordprofiler/. An English teacher could input wordlists that meets his own needs into it. For instance, Sport English word lists could be input beforehand to offer statistics of the text you pasted in the box.

6. Conclusion

This essay underlined the importance and usefulness of frequency for corpus research when it comes to categorizing vocabulary and creating lexical profiles of individual texts. The writer discussed how information about the occurrence of words can be usefully employed by language teachers and Sports English materials developers. Thanks to that many corpora nowadays can be accessed free of charge and are accompanied by resources on how to apply them to language teaching. Teachers’ awareness of benefits of corpus-informed teaching and ability of employing corpus techniques should be raised. English teachers need to acquire corpus literacy to interpreting corpus data and using which to produce teaching materials. It is hoped that with the continuing growth of corpus linguistics, more and more teachers, material developers and language teaching practitioners will start introducing element of this methodology and corpus-based tools into their everyday work. Meanwhile, more quantitative and qualitative researches are anticipated.

Acknowledgment

The research for this paper is financially supported by Shandong Sports Institute, “The Establishment and Use of Multimodal Micro-corpus of Sports English”, grant No.20151201.
References

[1] Cortes, V. (2013). Corpora in language for specific purposes research, in C. A. Chapelle (ed.) The Encyclopedia of Applied Linguistics. New York: Blackwell Publishing.

[2] Hyland, K. (2007). English for specific purposes: Some influences and impacts, in J. Cummins and C. Davison (eds.) International Handbook of English Language Teaching. New York: Springer, 391–402.

[3] Hewings, M. (2012) Using corpora in research, teaching, and materials design for ESP: an evaluation. Taiwan International ESP Journal 4/1: 1–24.

[4] Gilmore, A. (2015). Research into practice: The influence of discourse studies on language descriptions and task design in published ELT materials. Language Teaching, 48, 4, 506–530.

[5] Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design, in S. Granger, G. Gilquin and F. Meunier (eds.) The Cambridge Handbook of Learner Corpus Research. Cambridge: Cambridge University Press, 485–510.

[6] Nation, I. S. P. (2011). Research into practice: Vocabulary. Language Teaching, 44, 4, 529–539.

[7] O’Keeffe, A., McCarthy, M. and Carter, R. (2007). From Corpus to Classroom. Cambridge: Cambridge University Press

[8] Dang, T. N. Y. and Webb, S. (2016). Evaluating lists of high-frequency words. International Journal of Applied Linguistics, 167, 2, 132–158.

[9] Schmitt, N. and Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. Language Teaching, 47, 4, 484–503.

[10] Cobb, T. and Horst, M. (2015). Learner corpora and lexis, in S. Granger, G. Gilquin and F. Meunier (eds.) The Cambridge Handbook of Learner Corpus Linguistics. Cambridge: Cambridge University Press, 185–206.