Philip S. Riccobono*

Kobe University, Japan
riccobono@people.kobe-u.ac.jp

TRIANGULATING DIAMOND TALK: IDENTIFYING TECHNICAL SPOKEN VOCABULARY IN ENGLISH FOR BASEBALL PURPOSES

Abstract

Traditionally considered the all-American sport, baseball has progressively been internationalized in recent years. Non-native English-speaking players comprised 29.8% of the 2017 opening-day Major League Baseball team rosters, representing a record 19 nations and territories (MLB, 2017). In 2012, international players filled 3,382 spots on team rosters at the minor league level (MLB, 2012). Additionally, English is the *lingua franca* for premiere international baseball events. To address the increasing globalization of baseball, a new subtype of English for Occupational Purposes, namely, English for Baseball Purposes, is needed in order to teach and learn the technical vocabulary essential for communicating within this discourse community (Coxhead, 2013; Nation, 2012). This study reports on the construction of the Baseball English Corpus (BECO), and offers specialized vocabulary sets based on this corpus and the recommendations of ethnographic interview participants who are core users or stakeholders of Baseball English. By utilizing a mixed-methodology design to finalize the technical spoken baseball word and phrase lists (Chung & Nation, 2003; Tangpijaikul, 2014), this study also provides insights into various methods for identifying a technical corpus based lexicon as well as some pedagogical implications for Baseball English teachers and learners.

Key words

English for Baseball Purposes, Baseball English Corpus, tertiary sport, corpus linguistics, technical vocabulary.

* Corresponding address: Philip S. Riccobono, 3-15-41, Asahigaoka, Tarumi-ku, Kobe, Hyogo, 655-0033, Japan.

114

1. INTRODUCTION

Baseball has become increasingly internationalized in recent decades. Non-native English-speaking (NNES) baseball players comprised 29.8% of the 2017 opening-day Major League Baseball (MLB) team rosters, representing a record 19 nations and territories (MLB, 2017). In 2012, international players filled 3,382 spots on team rosters at the minor league level (MLB, 2012). Of the multinational talent at the highest echelons of professional baseball in North America, that is, MLB, only nine players born outside the U.S. hailed from English-speaking countries (MLB, 2014, 2015, 2016). Because English is not the L1 for most baseball specialists born outside the U.S., a significant and growing need exists for them to attain a certain level of proficiency in the English used in baseball, especially the technical baseball lexicon, to support their professional competencies. Additionally, English is the *lingua franca* in many high-level international baseball events (e.g. the Olympics, Premier 12, and World Baseball Classic).

In specific cases, professional NNES baseball players who moved to MLB, the world's largest and most prestigious professional baseball league based in the United States and Canada, attributed a part of the responsibility for their unsuccessful performances to their lack of Baseball English proficiency (Kelley, 2009; Park, 2016; Pennington, 2011; Shipgel, 2005). Furthermore, empirical data suggest a need for lexical improvement among NNES baseball players and specialists. Riccobono's (2018) Technical Baseball Lexical Assessment, which measures aural knowledge of the technical baseball lexicon, indicated a need for NNES baseball specialists to acquire baseball vocabulary.

To the best of my knowledge, however, studies in the corpus-building literature have not extracted baseball vocabulary. Therefore, in comprehensive research, there exists a need to facilitate Baseball English training for tertiary learners at university physical education courses, baseball teams, and professional or amateur NNES baseball specialists. The vocabulary examined in this study originates from the Baseball English Corpus (BECO), an ethnographically predisposed corpus involving spoken and written English discourse recommended by the core users (CU) of Baseball English, inclusive of those exercising baseball industry discourse, such as baseball coaches, players, umpires, team front-office personnel, and scouts. The other half of BECO involves peripheral users of Baseball English texts, i.e. individuals who do not work within organized baseball, such as media writers, play-by-play announcers, podcasters, and bloggers. Increasing a sport's vocabulary could promote better communication and effective action in relation to involvement in a particular sport (Stolz & Pill, 2016). Given the current circumstances, this research provides a long-needed starting point for the establishment of English for Baseball Purposes (EBP) vocabulary. The results may

benefit teachers and students of physical education courses at the university level or English for Sport pedagogy, in addition to baseball players, umpires, coaches, Baseball English core users, and fans.

The remainder of this paper is structured as follows: section 2 reviews literature which forms the theoretical background for the need for identifying technical or specialized vocabulary, while section 3 provides an account of corpus-based and intuited approaches to identifying technical baseball vocabulary. Section 4 includes the study's results and discussion whereas section 5 offers pedagogical implications for Baseball English learners and practitioners. The final section includes some concluding remarks.

2. LITERATURE REVIEW

2.1. Technical vocabulary

Technical words represent vocabulary that is used very frequently in a specific text or specialized domain but infrequently or not at all in other fields (Chung & Nation, 2003, 2004; Coxhead & Hirsh, 2007; Wang, Liang, & Ge, 2008). Although it is not possible to calculate exactly the total number of technical words used in a discipline, Coxhead and Nation (2001: 252) suggest that the "technical vocabulary of a discipline accounts for probably 1,000 words or less." Instances of technical words include acronyms, abbreviations, formulas, symbols, and associated field-specific words (Nation, 2001). Chung and Nation (2003), for example, identified that the field of anatomy has a large technical vocabulary comprising 4,270 word types that account for 37.6% of all word types in the corpus, while applied linguistics texts have a relatively small technical vocabulary comprising 835 technical word types that account for a considerably smaller proportion (16.3%) of all word types in the corpus. Coxhead and Demecheleer's (2018) Plumbing word list covers more than 30% of the written corpus and slightly more than 11% of the spoken corpus. However, the methodology underlying these findings merits consideration when identifying technical lexicon, for instance, corpora balance and representativeness, extraction measures, lemmatization, and inter-rater size (Riccobono, 2018). As shown in this study, the distinctions between technical and non-technical vocabulary are real and relevant and are, perhaps, most effectively defined using a corpus and semantic rater scale approach, which is described in the methodology section of this paper.

116

2.2. Vocabulary lists and corpus-based ESP and EOP vocabulary studies

Vocabulary lists may address the fundamental need of a beginner-level learner to better communicate in a field-specific discourse (e.g. Coxhead, 2013; Lewis, 1993, 1997, 2000; Nation, 1993; Tangpijaikul, 2014), such as Baseball English. However, learning to communicate in a specific field greatly relies on vocabulary size (Chung & Nation, 2003; Hutchinson & Waters, 1987), and a rigorous and strategic approach to drive linguistic comprehension and functional proficiency is necessary. Identification of the deficiencies in learning English vocabulary requires an understanding of not only a learner's lack of knowledge of words, collocations, and extended word sequences (N-grams), which is linked to the learner's specific field of study or occupation (Martínez, Beck, & Panza, 2009; Wang et al., 2008; Ward, 2009), but also a lack of semantic and syntactic comprehension, which leads to insufficiencies in L2 discursive reasoning, analysis, and application.

Given the role of English as a *lingua franca* in baseball, NNES baseball specialists working in an English-only environment currently lack access to empirical research data for learning the technical spoken baseball lexicon of statistical keyness. Instead, they rely on baseball glossaries that consist of intuited vocabulary and are geared principally toward a homogenous group of readers whose unique language groups are ill-defined or unaddressed (Riccobono, 2018).

117

Though intuited vocabulary lists provide value, corpus-based findings could strengthen a list based on empirical findings (Braun, 2005; Tomlinson, 2012). Thus, the present study is aligned with other studies on corpus-based English for Specific Purposes (ESP) and English for Occupational Purposes (EOP) vocabulary that have aimed to provide authentic, technical or essential lexicon. For instance, in Hirata's (2019) study, the cover letter wordlist was prepared to address the writing of a cover letter as an essential part of the job application process for tertiary learners and English language learners.

Similar corpus-based research, which identifies specialized vocabulary, highlights the possibility of creating a specialized list of words for specific texts depending on the need of the curriculum with no limits on frequency or general usage (Hunston, 2002; Nation, 2001). For example, Tangpijaikul (2014) generated essential words for English for Business students in Thailand by analyzing business newspapers and selecting the words that appeared frequently in them. Words such as *broker*, *executive*, and *surplus* are not found in either West's (1953) General Service List (GSL) or Coxhead's (2000) Academic Word List (AWL), but they are ubiquitous and very important in the texts used in the business world. However, in line with Coxhead's (2013) AWL and West's (1953) GSL, words or common core

approaches to vocabulary will remain as candidates of technical words under consideration for inclusion in the wordlist prepared in this study.

Chung and Nation's (2003, 2004) search for technical anatomy words provides a useful approach to identifying technical vocabulary by employing a rater scale approach, which is adopted in this study as well. Additional research categorizing specialized vocabulary includes Wang et al.'s (2008) study, Chen and Ge's (2007) and Hsu's (2013) works on the words used in medical research; Hou's (2014) wine corpus-based study to identify a wordlist of written keywords, and Martínez et al.'s (2009) efforts to document salient agricultural academic words. Other scholars have researched rudimentary English engineering terms that can benefit undergraduate engineering majors (Mudraya, 2006; Ward, 2009). Vongpumivitch, Huang, and Chang (2009) sought non-academic content words used in applied linguistics research papers, and Khamphairoh and Tangpijaikul (2012) explored technical vocabulary in insurance research articles.

In a more recent study of field-specific vocabulary, Coxhead and Demecheleer (2018) utilize both spoken and written corpora to identify little known technical plumbing terms. Although plumbing discourse for L1 and L2 stakeholders may employ written and spoken registers, baseball specialists are more likely to initially focus on the spoken lexicon, as will be revealed in ethnographic interviews later in this paper. Notably, Coxhead and Demecheleer (2018) do not indicate which lexicon – spoken or written – L1 and L2 plumbers need to focus their initial efforts on, which represents a gap in the literature. As discussed later, core user spoken discourse is identified as the representative register of Baseball English, thus impelling this study to examine such lexicon.

118

In one of the more significant studies in this arena, Nelson (2006) used a corpus-based approach to investigate Business English by examining both written and spoken data. Keywords that appeared more significant and more frequent in Business English than in general English were examined. However, Nelson (2006) did not provide a system for determining a metric to define the technicality of such business keywords, as Tangpijaikul (2014) did by using a semantic rater scale.

Corpus-based studies on specialized vocabulary are rare in sport in general and, to the best of my knowledge, non-existent in baseball in particular. However, a study examining the vocabulary used in the sport of body-building was identified (Murray, 1984). Even so, the methodology used in that study was inconsistent with the corpus approach; instead, it was closer to the dictionary approach. In addition, opinions vary on which method(s) researchers ought to implement when compiling vocabulary lists. Some scholars have researched extended word sequences (also called N-grams, lexical units, phrases, lexical phrases, and lexical chunks) in ESP/EOP, for instance, in dentistry (Pinna, 2007); in business, (Camiciottoli, 2007; Nelson, 2000), and in aviation (Aiguo, 2007). However, except for Martinez and Schmitt (2012), who used a dual approach-computer and manual

vetting-, none of these studies provided technical N-grams by employing a corpus-based approach, especially not in conjunction with raters using a semantic scale. Therefore, through the present study, an attempt is made to fill this gap.

2.3. Approaches for determining technical single-unit word types

The identification of technical words in a corpus is often difficult (Chung & Nation, 2003, 2004; Pearson, 1998). As stated by Cabré Castellví (1999), Bowker and Pearson (2002), and Chung and Nation (2003), some of the identification procedures draw on statistical formulas and quantitative data, such as comparing the frequency and range of words in a specialized corpus versus those in a general or a discipline-specific corpus. These comparisons are made using term extraction procedures. However, frequency criteria may lead to ambiguous findings insofar as the procedure may overlook the actual meanings that words acquire in different contexts. As Pearson (1998) states, the 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 Castellví (1999) and Coxhead and Nation (2001) pointed out, frequency counts typically reveal that many of the topic-related words in a corpus are actually general words that have a more specialized meaning in particular fields, or alternatively, they are words that are borrowed from other disciplines and applied in specialized ways to the new corpus. These claims highlight that technical words include not only highly specialized words, which have a single meaning and occur frequently in a particular discipline, but also words that are formally similar to and used as general words or other discipline-specific words. Thus, technical words are often semantically elastic because they encompass signification in general language, as well as specialized meanings in one or more specific disciplines. This applies to Baseball English as well, and therefore, as reported in this study, the words collected in a corpus for this specialized field must be examined by field experts.

119

Except for Tangpijaikul (2014) and Chung and Nation (2003, 2004), the aforementioned studies (on the identification of technical vocabulary) did not indicate the application of a semantic rating scale to the extracted specialized corpus-based vocabulary. The lack of a mixed-method design in constructing wordlists was considered when choosing an approach that would raise the validity of vocabulary lists in this study, as will be discussed further in the following section.

Therefore, the present work aims to fill the research gap in corpus-based EOP vocabulary studies by providing field-specific single-word units and N-grams for NNES baseball specialists who are already involved in or are considering participating in an English-speaking baseball environment, as well as for learners

and teachers of Baseball English, to enhance their Baseball English proficiency. The following research question guides this paper: *Which corpus-based single-word units and N-grams belong in a technical core-user spoken baseball vocabulary lists (appropriate for learners of Baseball English)?*

3. METHODOLOGY

3.1. Ethnographic interviews

To justify the need for a study on EBP and the accompanying vocabulary, a triangulated approach involving ethnographic uniform interview questions posed to Baseball English stakeholders (*N=6*) was employed in this study. The Baseball English stakeholders included tertiary coaches and former professional and amateur and tertiary players, meeting all of Davies's (2003) criteria for identifying native speakers. All but one of ethnographic interviews were conducted via telephone, Skype and Facetime due to participants residing in South Korea, Japan and the United States, with the researcher based in Japan. One interview was conducted face-to-face in Tokyo, Japan.

120

During interviews, the researcher asked the interviewees questions from a 10-item survey related to Baseball English. This instrument was based on Hong and Jhang's (2010) use of ethnographic interviews when creating their Maritime English Corpus and was adapted by the researcher to meet the needs of this study. The initial section of the questionnaire first and foremost established whether a need existed to explore, teach, and learn Baseball English, the core of EBP. Once research exigency had been established, the survey then investigated which register and genres required the greatest amount of attention for NNES baseball specialists and learners of Baseball English. Qualitative-focused coding (Saldana, 2009) was used to identify the recurring themes in the interviews (Charmaz, 2006; Saldana, 2009), which included the need for NNES baseball specialists and Baseball English learners to have baseball vocabulary knowledge from the essential core user spoken lexicon utilized by stakeholders. Most opined that spoken lexicon was of more value to NNES Baseball English over the written one since baseball players spend more time on the field, listening to specialized baseball receptive lexicon (especially in international baseball competitions where English represents the *lingua franca*) as opposed to reading written documents such as contracts, scouting reports, tutorials. However, interviewees agreed that the written core user lexicon is of value, only in the later phases of the Baseball English learning process.

As for recommendations toward the content of BECO from the six stakeholders, the author took these into account and they will be discussed in the next section.

3.2. Compiling the BECO corpus

The methodology used to compile the in-house corpus BECO was based on the suggestions by Coxhead (2000) and Biber (2003, 2009), in addition to those of Tangpijaikul (2014) for specialized corpus design. As Coxhead (2000) stated, collecting various short, medium, and long texts increases the representativeness of the corpus and decreases bias. The corpus size should be approximately 200,000 words to reach acceptance requirements, and the texts should be collected from diverse sources with a standard reference database where "information can be tagged as to sources of the written materials or texts" (Hong & Jhang, 2010: 974). Hence, this study categorically built a standard reference database categorized by file, description, genre (code), word count, time (when applicable, available), and (text) source (Biber & Conrad, 2009; Hong & Jhang, 2010; Wynne, 2004). All BECO texts conformed to those parameters.

121

All genres included in BECO were suggested by the six ethnographic participants in this study. The BECO CU spoken subcorpus texts include the utterances of coaches, players, scouts, and umpires during baseball activities, including practice sessions, games, private training, meetings, clubhouse/locker room conversations, news conferences, media interviews, instructional tutorials, health and wellness seminars, and motivational speeches from sports psychologists. CU written texts included contracts/collective bargaining agreements, tutorials, scouting reports or player evaluations: written reports assessing a baseball player's skills, rules, coaching guides, playbooks, CU editorials, book excerpts. The peripheral user spoken texts consisted of the following genres: play by play game announcers, fans discussing baseball, TV and radio baseball talk shows, radio and game broadcasts, podcasts. Finally, peripheral user written texts consisted of newspapers, books, blogs, baseball media guides, scouting reports, book excerpts, online chats, analytics research, Wikipedia articles, children's books, educational materials, adolescent literature: fiction and non-fiction.

Therefore, BECO contains 813,921 tokens, including 21, 548 word types from spoken and written registers across various genres of Baseball English (both core and peripheral users) (Table 1). Given that this study prioritizes Baseball English learners, as informed by the six Baseball English CUs in the ethnographic interviews, the focus is on identifying spoken vocabulary used by the CUs.

Register/Subcorpora	Number of Texts	Number of Word Tokens	Word Types
Core-user spoken	349	211,310	7,134
Core-user written	94	200,355	9,462
Peripheral user spoken	68	201,401	8,847
Peripheral user written	80	200,889	13,797
	Total number of texts in corpus	Total number of word tokens in corpus	Total word types in corpus
BECO **Total**	**591**	***813,921**	***21,548**

*Overlap exists across BECO subcorpora types.

Table 1. Composition of the Baseball English Corpus (BECO)

In a comparison of the top 100 ranked statistically significant key words and trigrams (Bednarek, 2015) of all four BECO subcorpora, CU spoken emerged as the discourse community with the most unique lexicon, exhibiting the least overlap with the other subcorpora, namely CU written, peripheral user spoken, and peripheral user written. As can be inferred from Figure 1, the CU spoken subcorpus is balanced and representative of this subfield, with sufficient texts and tokens in each sub-area. Therefore, the CU spoken subcorpus served as the subcorpus of analysis in this study.

Genre	# of words	Files
Game Discourse	39733	123
Practice Discourse	23192	39
Pre/Post Game Discourse	19317	55
Media Discourse	28423	32
Instructional Pitching	27991	33
Instructional Hitting and Running	19041	22
Instructional Position Players:	16033	23
Health and Wellness	26270	22
Total	211310	349

Figure 1. Compilation of CU spoken subcorpus

3.3. Identifying vocabulary of keyness

A fundamental characteristic of the act of compiling and analyzing corpora is the use of machine- or computer-readable formats (Baker, 2006; Bednarek, 2015;

Hong & Jhang, 2010). This study employed Anthony's (2016a) AntConc 3.5.0 to extract single-word units and Drouin's (2010) TermoStat Web 3.0 to extract N-grams. The researcher compared reference corpora through log-likelihood exams to identify the vocabulary of statistical keyness or the key vocabulary (Culpeper, 2009; Grabowski, 2015; Scott, 2010; Tangpijaikul, 2014) in the baseball subcorpus.

A word or phrase is considered *key* or of *keyness* value only when its frequency rank in the target (sub)corpus (CU spoken subcorpus of BECO in this study) is high compared to its rank in the reference corpus. Studies have shown that the reference corpus comparison approach is a suitable starting point for numerous corpus-based vocabulary analyses (Evison, 2010; Tangpijaikul, 2014). Others have suggested that advanced keyness analysis of the lexicon is the most effective approach to identifying technical vocabulary (Baker, 2006; Hunston, 2002). Additionally, Mudraya (2006) argued that keywords provide a better sense of technical words because the most frequent words in a specialized corpus are, in fact, sub-technical and nontechnical. Thus, the keyness approach goes beyond mere frequency counts because it compares two frequency lists by conducting a statistical comparative analysis of a target corpus. Moreover, a substantial and growing body of research has confirmed the benefits of keyword/phrase lists, including their robustness, which influenced their use in this study in several capacities, for instance, as a jumping off point (Chung & Nation, 2003; Hunston, 2002; Kwary, 2011; Tangpijaikul, 2014) to determine technical vocabulary before it is determined by field stakeholders using a semantic rating scale as an instrument.

123

This study employed the Corpus of Contemporary American English (COCA) spoken texts (for single-word units) and TermoStat's English Corpus, which consists of the British National Corpus and newspaper articles from various North American publications. TermoStat, a web-based platform, devises its own preset reference corpus. COCA represents American English, which encapsulates the core of the spoken Baseball English used in this study. Ideally, the use of COCA spoken as a reference corpus for extracting key N-grams would be preferred, but AntConc did not offer the functionality to extract key N-grams, whereas TermoStat did. The TermoStat reference corpus was built into the application and was unmodifiable. Figure 2 demonstrates how the approach used to identify key words/N-grams was combined with a lexical profiling approach to identify technical vocabulary and an expert intuition-based rating scale approach to identify technical keywords/N-grams. The keyness value cut-off point (for additional analysis) was 30 for single-word units and 9 for N-grams, whereas Tangpijaikul (2014) set a threshold of 160 for the Business English wordlist, though he later realized that a higher threshold eliminated potential technical words.

3.4. Lemmatization

Words of the same base but with different inflectional affixes (e.g., *run*, *runs*, *runner*, *runners*) were subjected to lemmatization as one lexical item and considered to be members of the same word family at Level 2 in Bauer and Nation's (1993) classification of word families: "Regularly inflected words are part of the same family. The inflectional categories are plural, third person singular present tense, past tense, past participle, *-ing*, comparative, superlative, possessive" (Bauer & Nation, 1993: 270). Thus, lemmatization served to economize the lexis while creating consistency across the Technical Spoken Baseball Wordlist. Therefore, by using AntConc, keywords were lemmatized (Chung & Nation, 2003, 2004; Tangpijaikul, 2014).

3.5. Five-step procedure for identifying technical baseball terms

Technical words were identified according to the following five-step procedure (Figure 2):

(1) Frequency lists were formed.

(2) All running words in the BECO CU subcorpus were identified for their keyness values based on a comparison of the words to a reference corpus, resulting in a set of keywords (Tangpijaikul, 2014).

(3) Keywords were lemmatized (Tangpijaikul, 2014), but N-grams were not.

(4) Proper names and abbreviations or acronyms were excluded in line with Chung and Nation (2003) and Tangpijaikul (2014); these studies suggest that it is worthwhile to exclude names and places from the keyword/N-gram list.[1]

(5) Keywords/N-grams that remained on a semantic scale were rated, ranging from words and N-grams whose meanings were related to the field of baseball to those having no semantic relationship with baseball activities.

124

---

[1] However, following Chung and Nation (2003), this study retained a few abbreviations as candidates for raters. Furthermore, the researcher filtered out abbreviations consisting of proper names (Chung & Nation, 2003, 2004; Tangpijaikul, 2014).
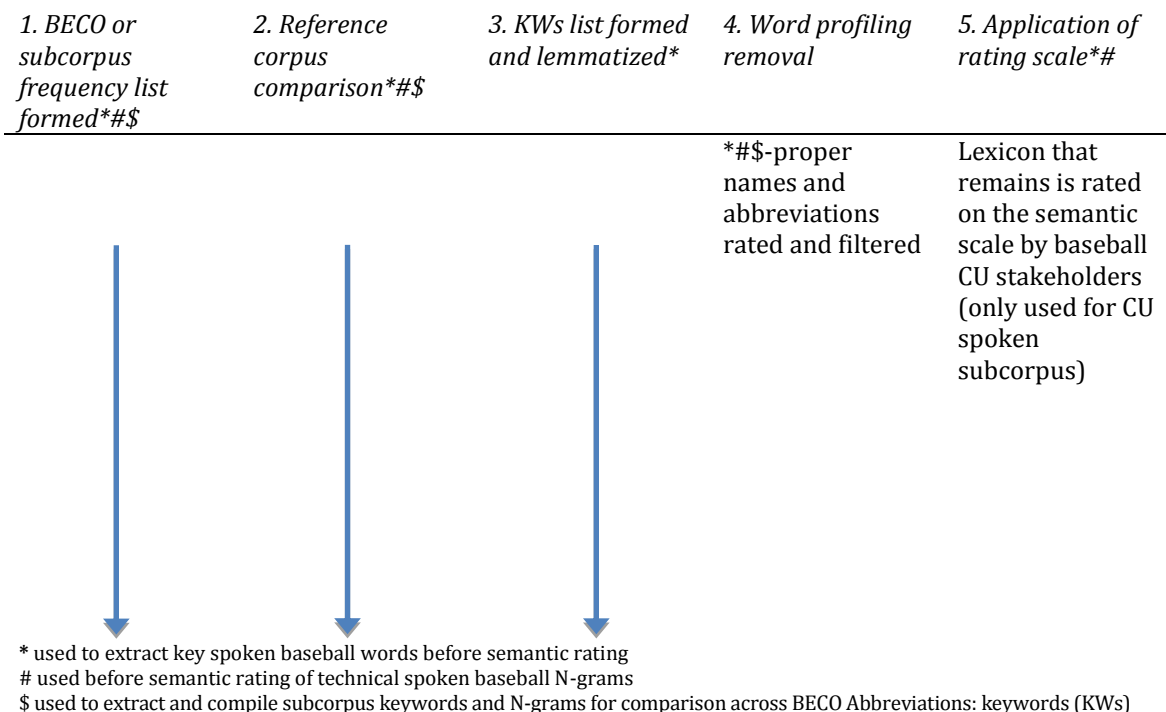
1. BECO or subcorpus frequency list formed*#$	2. Reference corpus comparison*#$	3. KWs list formed and lemmatized*	4. Word profiling removal	5. Application of rating scale*#
			*#$-proper names and abbreviations rated and filtered	Lexicon that remains is rated on the semantic scale by baseball CU stakeholders (only used for CU spoken subcorpus)

* used to extract key spoken baseball words before semantic rating
\# used before semantic rating of technical spoken baseball N-grams
$ used to extract and compile subcorpus keywords and N-grams for comparison across BECO Abbreviations: keywords (KWs)

Figure 2. Steps followed to extract technical spoken keywords/N-grams from BECO

125

3.6. Four-point semantic rater scale

After identifying the keywords/N-grams, the qualitative component for identifying the technical vocabulary was executed using a four-point rater scale (Appendix 1). The scale was used to generate semantic ratings of the baseball lexicon by individuals with extensive knowledge of the subject area (Chung & Nation, 2003, 2004). Chung and Nation (2003, 2004) and Tangpijaikul (2014) recommend that professionals with experience in the given field should be recruited as raters. To ensure reliable intuition when rating the lexicon, the researcher recruited five raters who worked in the field of baseball – Baseball English core users – to identify single-word units and N-grams. The raters were baseball coaches and former players, and they were trained for the purpose of this study based on previously recommended guidelines (Chung & Nation, 2003; Tangpijaikul, 2014).

The degree of agreement among the five raters at each step of the scale was evaluated to find any tendencies toward bias at any step (Chung & Nation, 2003; Tangpijaikul, 2014). As cited in Chung and Nation (2003), a raw accuracy score of 0.7 constitutes a desirable reliability threshold for rating items according to the four groups or levels of the semantic scale. In the present study, each word or

phrase included in the final spoken baseball technical word/phrase list had a reliability rating of 0.8 for any combination of Step 3 or 4 (among raters). That is, the lexicon rated and agreed upon by at least 80% of the five raters was retained as the final list of technical spoken baseball words/N-grams.

4. RESULTS AND DISCUSSION

4.1. Constitution of technical vocabulary lists

The Technical Spoken Baseball Wordlist (Appendix 2) consists of 2.4% technical words (169 single-word unit types) and 7,134 word types from the target CU spoken corpus, while the technical spoken baseball phrase list includes 352 CU baseball N-grams (Appendix 3). These lists were derived from the five raters' judgments about the keyword and N-gram lists, which included 424 single-word units and 779 N-grams, respectively, after filtration (Figure 2 above, steps 1–4). The single-word unit list size conforms to Coxhead and Nation's (2001) view that a technical vocabulary likely includes 1,000 words or fewer per discipline, and the wordlist consisting of 169 word types in this study represents a distinct niche. Additionally, the text coverage of the technical words in this study (2.4%) is less than that in aforementioned technical vocabulary studies (ranging from 16% to 38% text coverage) on anatomy (Chung & Nation, 2003), applied linguistics (Chung & Nation, 2003), cover letters (Hirata, 2019), and plumbing (Coxhead and Demecheleer, 2018). Notably, these coverages are for written texts. The findings of the present study suggest a lower text coverage of spoken technical vocabulary as opposed to the written one. The wordlist obtained in this study is more in line with Coxhead and Demecheleer's (2018) Plumbing wordlist coverage of just over 11% compared to the corresponding spoken corpus. Further discussion must consider the differences in the methodologies used in each study, which perhaps affect the text coverage of wordlists.

126

4.2. Inter-rater reliability

The Technical Spoken Baseball Wordlist demonstrates strong inter-rater reliability, indicating that the four-point semantic rating scale yields viable lists. By using this scale to choose the technical single-word units when rating all the 424 keyword list types, which are precursors to the final technical single-unit wordlist, the raters achieved a high degree of inter-rater reliability. The average Intraclass Correlation Coefficient (ICC) measure was .875, with a 95% confidence interval ranging from

.845 to .899, $F(423, 1692) = 8.992$, $p < .001$. This finding demonstrates a higher inter-rater reliability benchmark than the .70 suggested by Chung and Nation (2003). Therefore, the Technical Spoken Baseball Wordlist is a reliable triangulated list for Baseball English learners. Similarly, when performing keyness ratings on a four-point scale for all 779 N-grams generated from the technical spoken baseball key N-grams list by using TermoStat Web 3.0, the raters achieved a high degree of inter-rater reliability. The average ICC measure was .887, with a 95% confidence interval ranging from .859 to .908, $F(778, 3112) = 10.258$, $p < .001$. Finally, such constructs – Technical Spoken Baseball Wordlist and Technical Spoken Baseball Phrase list – can perhaps serve as foundations for the development of lexical-based Baseball English pedagogy.

4.3. Lexical profiling

Anthony's (2016b) AntWordProfiler 1.4.1m application was used to generate a breakdown of GSL, AWL, and off-list words across the Technical Spoken Baseball Wordlist. The author retained the words (as candidates for semantic raters to analyze) found in GSL and AWL, as well as those in abbreviations and proper nouns, reaching step 3 or 4 on the rater scale. If the author were to filter out all GSL and AWL types from the keyword list following Tangpijaikul's (2014) methodology, pertinent baseball lexicon (e.g. *ball*, *base*, and *runner* from GSL 1; *bag*, *hitter*, and *slide* from GSL 2; *consistency*, *fundamentals*, and *target* from AWL) would have been omitted from the Technical Spoken Baseball Wordlist (Figure 3). However, these words were included in the final Technical Spoken Baseball Wordlist based on the raters' validation (and agreement) with the use of a semantic scale. This finding confirms that technical vocabulary often overlaps with GSL and AWL types (Chung & Nation, 2003; Coxhead & Hirsh, 2007; Sutarsyah, Nation, & Kennedy, 1994). Nearly 60% of the Technical Spoken Baseball Wordlist consists of off-list words that are not included in the GSL or AWL (Figure 3). This finding perhaps increases the value of the list, which consists of common low-frequency and unknown words in general English. Moreover, the single-word units list contained more than 63% of AWL and off-list word types, which indicated a higher than typical lexical density (Stubbs, 2001). This is in line with other studies' specialized findings of keyword lexical density, for example, Hou's (2014) study related to a wine wordlist (66%). However, in Hou's study, the researcher did not use a semantic rater scale to identify technical words. Finally, this study may have a higher lexical density when employing Hou's (2014) (less robust) metric for identifying technical words.

127
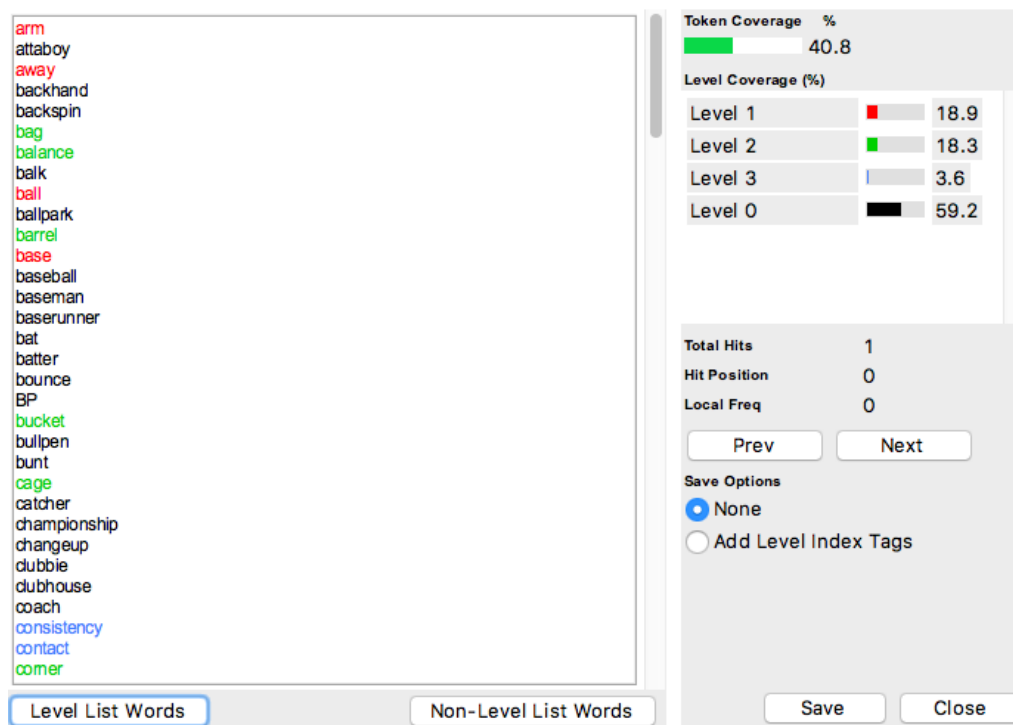
Note: Level 1: GSL 1 (29 of 1,000 types), Level 2: GSL 2 (29 of 1,000 types), Level 3: AWL (6 of 570 types), and Level 0: off-list or technical words from core-user spoken (100 types).

Figure 3. Screenshot from AntWordProfiler 1.4.1m showing the distribution of 169 technical spoken baseball words across GSL 1 (red), GSL 2 (green), AWL (blue), and off-list (black).

4.4. Conspicuous non-technical baseball vocabulary

Although certain vocabulary types were not considered technical by the five inter-raters, these types represented the lexicon of statistical keyness, scientifically proven to appear uniquely across spoken Baseball English (Table 2 in red). For example, *fuck* or *fucking* is often associated with anger, aggression, and hostility, and/or the intent to threaten, insult, or demean, or as a modifier (Beers-Fägersten, 2007; Hobbs, 2013). Half of the ethnographic participants in the study echoed that foul language or cussing is a pliable part of the CU lexicon. Specifically, *fuck*, *shit*, and *motherfucker* (in varying orders) have been rated as the most offensive (Beers-Fägersten, 2007; Jay, 2009). Therefore, these single-word units and N-grams of keyness extracted from the BECO merit reference for their role in North American baseball discourse (Table 2). These types were not considered technical by the stakeholders, albeit a gap exists between the corpus and the intuited baseball

lexicon, for instance, *bring me the cheese* and *dotting a gnat's ass.* Therefore, a combined CU-intuited and corpus-driven list may be examined in a future study related to EBP.

Classification	Examples of spoken core-user statistically significant key vocabulary aligned with stakeholders' recommendations
Technical	**fungo, spin, oppo, forkball, inning, second base, center field, front side closure, second baseman double play depth position, split finger fastball, glove side**
Off-putting	oh shit, horseshit, fucking, bullshit, fucking opponent, goddamn **game**, piece of shit
Slang	*dotting a gnat's ass, that guy just tittied that* **ball**
On-general	*bring me the cheese, brush back*

Note. Bold represents CU spoken BECO-extracted types rated as technical spoken Baseball English. Red signifies statistically significant keywords and N-grams from the BECO CU spoken subcorpus but not considered technical by the five raters. *Italics* represent examples of CU spoken Baseball English lexicon intuited by the ethnographic participants but not found in BECO.

Table 2. Classifications of spoken baseball vocabulary informed by *N=6* stakeholders

As is evident from Table 2, throughout the ethnographic interviews, all six stakeholders indicated that North American baseball mainly includes spoken technical Baseball English lexicon, representative of off-putting, filled with slang, cuss/curse words, and non-general words/N-grams, resulting in a language that is problematic to decipher, even for native English speakers involved professionally in baseball. The BECO vocabulary findings corroborate the stakeholders' opinions. Notably, the CU vocabulary features were corroborated owing to the triangulation: ethnographic interviews, corpus, and an intuition-based but structured semantic rater scale approach to identifying these terms. Consequently, there may exist a need for a separate nontechnical off-putting baseball vocabulary list, which would be essential for communicating within a focused discourse community (Bednarek, 2015). This can be achieved in a future study focused on off-putting baseball lexicon.

129

5. PEDAGOGICAL IMPLICATIONS

Practitioners of Baseball English should not simply leave learners with vocabulary that they feel compelled to have knowledge of. In working toward EBP, we must ensure that learners appropriately use technical terms belonging to different baseball genres in various phonological and orthographical activities.

Consequently, a growing body of research demonstrates the value of learning the language in chunks, which enables L2 learners to acquire specialized vocabulary efficiently (Lewis, 1997, 2000; Coxhead & Demecheleer, 2018). In other words, this technique optimizes learners' ability to recognize and to use single-word units, aiding learners' comprehension and deployment of lexical structures in the forms of clusters or chunks of words (Eggington & Cox, 2013; Martinez & Schmitt, 2012; Nation, 2012). The word and phrase lists obtained in this study facilitate the abovementioned approaches.

Nation (2001) revealed that the knowledge of spoken word forms consists of recognizing such a lexicon in an aural sense, while having the ability to use the word in spoken form to convey its meaning, which is at the other end of the receptive-productive spectrum. Before learners' transition to productive use of Baseball English, receptive practice of specialized baseball terms for learners of varying levels needs attention. To this end, exercises involving sentence samples in concordance lines consisting of single-word units and N-grams extracted from BECO (Figure 4) may facilitate receptive practice (Buck, 1992; Khamphairoh & Tangpijaikul, 2012; Kohn, 2001; Nation, 2001, 2008).

For example, concordance lines could be useful in scaffolding activities, whereas some words are known to pose a significant polysemic challenge. For example, in baseball, *'ball'* in general usage refers to a round object or a grand event, but, in this discipline-specific parlance, it refers to a type of a particular *pitch count* as in *balls and strikes*. Thus, the linguistic symbol, the word *ball*, may be invoked in general English or technical baseball, but the signification of the word varies widely, depending on the context and field of use.

130

i. Remember the Word/Guess the Meaning

dugout

home plate

seam

ii. Gap-filling: Watch the video and fill in the blanks with the words you hear.

Note: The underlined words are deleted; learners will write them in the blanks.

Coach 1: Want me to tell you a little secret about <u>baseball</u>? When a guy <u>throws</u> a <u>slider</u> and a <u>curve</u>, one of them sucks.
Coach 2: I know that. But we are not using it to be…
Coach 1: Just to show it, yeah.
Coach 2: We're not, we're not trying to be effective with it. We are trying to show it. Oh shit! He's got a fucking <u>curveball</u>…

iii. Collocation match-up: Match the two parts of these phrases.

1. _____ *c* on   a. fastball
2. _____ *a* in   b. game
3. _____ *b* day   c. runner

iv. Cluster challenge: Complete the collocations using these words: fastball/runner/game

1. Baseball is a *<u>game</u>* of failure, right?
2. You got a *<u>runner</u>* on first.
3. He might try to stick a *<u>fastball</u>* in on you with 2 strikes.

v. Skill practice: Complete each sentence using the given words below.

1. He's one of the best players in the _____, and he will probably continue to get better.
a. minute b. *game* c. team
2. The cutter is a combination of the _____ and the slider, and it runs away from the batter.
a. curveball b. fork c. *fastball*

Figure 4. Vocabulary exercises

This type of intra-baseball word, which Hsu (2013) defined as crypto-technical, has its own meaning within baseball. Crypto-technical is significant here because it illustrates that these word types have a specialized meaning in a particular field, such as *ball*, *strike*, *run*, *slide*, and *slider* in baseball. Therefore,

enabling concordance analysis can negotiate meaning or scaffold Baseball English lexicon through the following exercises.

6. CONCLUSION

The constellation of approaches used in this research have allowed the construction of technical baseball vocabulary lists with high reliability. By not relying on intuition or corpus findings alone, the results of this study offer insights into the use of a triangulated, mixed-methods approach to other areas of ESP or EOP for constructing vocabulary lists. Moreover, this study augmented Chung and Nation's (2003) semantic rater scale approach by employing it to identify technical N-grams. In developing EBP as a discrete, specialized professional field, the present study suggests a highly feasible, vocabulary-centric pedagogy for Baseball English learners. Such pedagogies should be established based on the strategic instruction of technical words and N-grams derived from the CU spoken subcorpus of BECO.

This study may fill a void in the teaching and learning of baseball vocabulary. It can possibly support physical education or English for Sport learners and practitioners at the tertiary level, as well as NNES baseball specialists who find themselves without an interpreter. Finally, the introduction of the technical CU spoken lexicon serves as a step toward opening up EBP to all Baseball English learners, from Little League players in Asia, Latin American college players in North America, NNES umpires, and even NNES baseball fans listening to MLB broadcasts worldwide.

Yet, this study has several limitations. First, the five inter-raters checked CU spoken keywords and N-grams from one of the four subcorpora in BECO. Owing to feasibility concerns and raters volunteering their time, they did not evaluate the key lexicon of the other three BECO subcorpora, which could have helped them to possibly identify additional technical baseball lexicon that would be beneficial to Baseball English learners and facilitators. Next, the researcher may add additional hitter/runner and position player spoken texts. However, the addition of spoken texts is quite labor intensive. Increased funding in the form of grants or institutional support can potentially remedy this issue by allowing the researcher to hire competent professional transcriptionists. Thus, this approach may improve the representativeness and balance of BECO CU for achieving an optimal coverage of Baseball English lexicon.

Future studies on this topic could consider the distinctions among the various registers of Baseball English. In addition, future directions for research based on findings from the present study can include employing aural lexical assessment constructs such as vocabulary-based listening and oral response examinations, measuring NNES baseball specialists baseball lexicon proficiency; gaps between

132

BECO lexicon of keyness and stakeholder intuited Baseball English word/phrase lists; qualitative findings of coded needs analysis interviews from NNES baseball specialists; and prosodic analysis of lexicon within discourse communities of the four BECO subcorpora.

Acknowledgement

References

Aiguo, W. (2007). Teaching aviation English in the Chinese context: Developing ESP theory in a non-English speaking country. *English for Specific Purposes*, *26*(1), 121-128. https://doi.org/10.1016/j.esp.2005.09.003

Anthony, L. (2016a). AntConc (Version 3.5) [Computer software]. Tokyo, Japan: Waseda University. Retrieved from http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2016b). AntWordProfiler (1.4.1m) [Computer software]. Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net

Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253-279. https://doi.org/10.1093/ijl/6.4.253

Bednarek, M. (2015). Corpus-assisted multimodal discourse analysis of television and film narratives. In P. Baker, & T. McEnery (Eds.), *Corpora and discourse studies*: *Integrating discourse and corpora* (pp. 63-87). Hampshire: Palgrave Macmillan. https://doi.org/10.1057/9781137431738_4

Beers-Fägersten, K. (2007). Swear word offensiveness. *Saarland Working Papers in Linguistics*, *1*, 14-37.

Biber, D. (2003). Variation among university spoken and written registers: A new multi-dimensional analysis. *Language and Computers*, *46*(1), 47-70. https://doi.org/10.1163/9789004334410_005

Biber, D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics*, *14*(3), 275-311. https://doi.org/10.1075/ijcl.14.3.08bib

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511814358

Bowker, L., & Pearson, J. (2002). *Working with specialized language: A practical guide to using corpora.* London: Routledge. https://doi.org/10.4324/9780203469255

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, *17*(1), 47-64. https://doi.org/10.1017/S0958344005000510

133

Buck, G. (1992). Listening comprehension: Construct validity and trait characteristics. *Language Learning*, *42*(3), 313-357. https://doi.org/10.1111/j.1467-1770.1992.tb01339.x

Cabré Castellví, M. T. (1999). *Terminology theory, methods, and applications*. Amsterdam: John Benjamins. https://doi.org/10.1075/tlrp.1

Camiciottoli, B. C. (2007). *The language of business studies lectures*. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.157

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis.* London: Sage.

Chen, Q., & Ge, G.-C. (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. http://dx.doi.org/10.1016/j.esp.2007.04.003

Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language*, *15*(2), 103-116.

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System*, *32*(2), 251-263. https://doi.org/10.1016/j.system.2003.11.008

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238. https://doi.org/10.2307/3587951

Coxhead, A. (2013). Vocabulary and ESP. In B. Paltridge, & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 115-132). Hoboken: John Wiley & Sons. https://doi.org/10.1002/9781118339855.ch6

Coxhead, A., & Demecheleer, M. (2018). Investigating the technical vocabulary of plumbing. *English for Specific Purposes*, *51*, 84-97. https://doi.org/10.1016/j.esp.2018.03.006

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliquée*, *XII*(2), 65-78. https://doi.org/10.3917/rfla.122.0065

Coxhead, A., & Nation, I. S. 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. https://doi.org/10.1017/CBO9781139524766.020

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*. *International Journal of Corpus Linguistics*, *14*(1), 29-59. https://doi.org/10.1075/ijcl.14.1.03cul

Davies, A. (2003). *The native speaker: Myth and reality*. Clevedon, UK: Multilingual Matters.

Drouin, P. (2010). TermoStat Web 3.0 [Computer Software]. Retrieved from http://olst.ling.umontreal.ca/~drouinp/termostat_web/

Eggington, W. G., & Cox, T. (2013). Using elicited oral response testing to determine the need for an interpreter. *Harvard Latino Law Review*, *16*(1), 127-146.

Evison, J. (2010). What are the basics of analysing a corpus? In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 122-135). London: Routledge.

Grabowski, L. (2015). Keywords and lexical bundles within English pharmaceutical discourse: A corpus-driven description. *English for Specific Purposes*, *38*(2), 23-33. https://doi.org/10.1016/j.esp.2014.10.004

Hirata, M. (2019). An English cover letter essential wordlist for second language learners. *The Language Teacher*, *43*(1), 3-8. Retrieved from https://jalt-publications.org/sites/default/files/pdf-article/43.1tlt-art1.pdf

134

Hobbs, P. (2013). Fuck as a metaphor for male sexual aggression. *Gender and Language*, *7*(2), 149-176. https://doi.org/10.1558/genl.v7i2.149

Hong, S. C., & Jhang, S. E. (2010). The compilation of a maritime English corpus for ESP learners. *Korean Journal of English Language and Linguistics*, *10*(4), 963-985.

Hou, H. I. (2014). Teaching specialized vocabulary by integrating a corpus-based approach: Implications for ESP course design at the university level. *English Language Teaching*, *7*(5), 26-37. https://doi.org/10.5539/elt.v7n5p26

Hsu, W. (2013). Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. *Language Teaching Research*, *17*(4), 454-484. https://doi.org/10.1177/1362168813494121

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524773

Hutchinson, T., & Waters, A. (1987). *English for specific purposes.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511733031

Jay, T. (2009). The utility and ubiquity of taboo words. *Perspectives on Psychological Science*, *4*(2), 153-161. https://doi.org/10.1111/j.1745-6924.2009.01115.x

Kelley, S. (2009, April 8). Mariners Kenji Johjima steps up behind the plate. *The Seattle Times.* Retrieved from http://www.seattletimes.com/sports/mariners-kenji-johjima-steps-up-behind-the-plate/

Khamphairoh, T., & Tangpijaikul, M. (2012). Collocations of keywords found in insurance research articles: A corpus-based analysis. *Humanities Journal*, *19*(2), 166-188. Retrieved from https://so04.tci-thaijo.org/index.php/abc/article/download/54124/44924/

Kohn, K. (2001). Developing multimedia CALL: The Telos language partner approach. *Computer Assisted Language Learning*, *14*(3-4), 251-267. https://doi.org/10.1076/call.14.3.251.5788

Kwary, D. A. (2011). A hybrid method for determining technical vocabulary. *System*, *39*(2), 175-185. https://doi.org/10.1016/j.system.2011.04.003

Lewis, M. (1993). *The lexical approach: The state of ELT and the way forward*. Hove, U.K.: Language Teaching Publications.

Lewis, M. (1997). Pedagogical implications of the lexical approach. In J. Coady, & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 255-270). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524643.018

Lewis, M. (2000). Learning in the lexical approach. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 155-185). Hove, U.K.: Language Teaching Publications.

Major League Baseball. (2012, April 5). *Opening day rosters feature 243 players born outside of the U.S.* Retrieved from http://m.mlb.com/news/article/27953108/opening-day-rosters-feature-243-players-born-outside-of-the-us/

Major League Baseball. (2014, April 1). *2014 Opening day rosters feature 224 players born outside the U.S.* Retrieved from http://m.mlb.com/news/article/70623418/2014-opening-day-rosters-feature-224-players-born-outside-the-us/

Major League Baseball. (2015, April 6). *Opening day rosters feature 230 players born outside the U.S.* Retrieved from http://m.mlb.com/news/article/116591920/opening-day-rosters-feature-230-players-born-outside-the-us/

135

Major League Baseball. (2016, April 4). *Opening day rosters feature 238 players born outside the U.S.* Retrieved from http://m.mlb.com/news/article/170436708/opening-day-rosters-feature-238-players-born-outside-the-us/

Major League Baseball. (2017, April 3). *Opening day rosters feature record 259 players born outside the U.S.* Retrieved from http://m.mlb.com/news/article/222084690/opening-day-rosters-feature-record-259-players-born-outside-the-us/

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, *33*(3), 299-320. https://doi.org/10.1093/applin/ams010

Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes*, *28*(3), 183-198. https://doi.org/10.1016/j.esp.2009.04.003

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes*, *25*(2), 235-256. https://doi.org/10.1016/j.esp.2005.05.002

Murray, T. E. (1984). The language of bodybuilding. *American Speech*, *59*(3), 195-206.

Nation, I. S. P. (1993). Vocabulary size, growth and use. In R. Schreuder, & B. Weltens (Eds.), *The bilingual lexis* (pp. 115-134). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/sibil.6.07nat

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nation, I. S. P. (2008). *Teaching vocabulary: Strategies and techniques.* Boston: Heinle Cengage Learning.

Nation, I. S. P. (2012). Teaching language in use: Vocabulary. In M. Eisenmann, & T. Summer (Eds.), *Basic issues in EFL teaching and learning* (pp. 103-116). Heidelberg: University Winter Verlag.

Nelson, M. (2000). *A corpus-based study of the lexis of business English and business English materials* (Unpublished doctoral dissertation). University of Manchester, Manchester, United Kingdom.

Nelson, M. (2006). Semantic associations in business English: A corpus-based analysis. *English for Specific Purposes*, *25*(2), 217-234. https://doi.org/10.1016/j.esp.2005.02.008

Park, S. (2016). *Major league baseball and Asian-born baseball players* [Executive summary]. Retrieved from https://www.slideshare.net/SangmiPark3/mlb-asian-born-players

Pearson, J. (1998). *Terms in context.* Amsterdam: John Benjamins. https://doi.org/10.1075/scl.1

Pennington, B. (2011, July 23). Kei Igawa: The lost Yankee. *The New York Times.* Retrieved from http://www.nytimes.com/2011/07/24/sports/baseball/kei-igawa-the-lost-yankee.html?mcubz=3

Pinna, A. (2007). Exploiting LSP corpora in the study of foreign languages. In D. Gálová (Ed.), *Languages for specific purposes: Searching for common solutions* (pp. 146-162). Newcastle, UK: Cambridge Scholars Publishing.

Riccobono, P. S. (2018). *Towards English for baseball purposes (EBP): A corpus-based investigation of contemporary vocabulary for non-native English speaking specialists and learners of baseball English* (Unpublished doctoral dissertation). Kyung Hee University, Seoul, South Korea.

Saldana, J. (2009). First cycle coding methods. In J. Saldana (Ed.), *The coding manual for qualitative researchers* (pp. 45-145)*.* Thousand Oaks, CA: Sage Publications.

Scott, M. (2010). What can corpus software do? In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 136-151). London: Routledge.

Shipgel, B. (2005, November 23). Fluent in baseball, if not English. *The New York Times.* Retrieved from http://www.nytimes.com/2005/11/23/sports/baseball/fluent-in-baseball-if-not-english.html?mcubz=3

Stolz, S. A., & Pill, S. (2016). Telling physical education teacher education tales through pedagogical case studies. *Sport, education and society*, *21*(6), 868-887. https://doi.org/10.1080/13573322.2014.962495

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell Publishers.

Sutarsyah, C., Nation, P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based case study. *RELC Journal*, *25*(2), 34-50. https://doi.org/10.1177/003368829402500203

Tangpijaikul, M. (2014). Preparing business vocabulary for the ESP classroom. *RELC Journal*, *45*(1), 51-65. https://doi.org/10.1177/0033688214522641

Tomlinson, B. (2012). Materials development for language learning and teaching. *Language Teaching*, *45*(2), 143-179. https://doi.org/10.1017/S0261444811000528

Vongpumivitch, V., Huang, J.-Y., & Chang, Y.-C. (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. https://doi.org/10.1016/j.esp.2008.08.003

Wang, J., Liang, S. L., & Ge, G. C. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, *27*(4), 442-458. https://doi.org/10.1016/j.esp.2008.05.003

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, *28*(3), 170-182. https://doi.org/10.1016/j.esp.2009.04.001

West, M. P. (1953). *A general service list of English words. With semantic frequencies and a supplementary word-list for the writing of popular science and technology* (Revised and enlarged edition). London: Longmans, Green & Company.

Wynne, M. (Ed.) (2004). *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books.

137

PHILIP S. RICCOBONO holds a Ph.D. in TESOL from Kyung Hee University in Seoul, South Korea. He teaches English language learners in the School of Languages and Communication (SOLAC) at Kobe University, Kobe, Japan. His research interest relates to corpus-driven-technical-vocabulary for ESP. He has published on topics associated with corpus linguistics, vocabulary acquisition and media-assisted language learning. Aside from lecturing, researching and publishing, you may find him evaluating talent at a baseball game around the globe.

Appendix 1

Four-point rating scale to identify technical baseball words and phrases
(adapted from Chung and Nation, 2003 and Tangpijaikul, 2014)

Level 1: Function words and phrases that have <u>no particular relationship</u> to baseball. Words: (e.g. *a, an, the, is, among, really, very, cannot, article*) Phrases: (e.g. *is going to, there in the, we are, why don't you, a lot of very, much of everything*)
Level 2: Words and phrases <u>marginally related to</u> baseball whose meanings may not be concerned directly with, but can be interpreted as being related to, baseball and its activities. Words: (e.g. *behind, below, middle, superior, body*) Phrases: (e.g. *big things, great guy, group of guys, speed differential*)
Level 3: Words and phrases <u>closely related to</u> baseball and its activities that can still be used in other fields. Such words and phrases are also used in general language. Words: (e.g. *slider, catcher, glove, mitt, scorecard, donut, barrel*) Phrases: (e.g. *bad throw, pro ball, short turn, big lead, nice start, locker room*)
Level 4: Words and phrases <u>specific</u> to baseball not likely to be known in general language. These words and phrases have clear restrictions of usage, depending on the subject field. Words: (e.g. *baserunner, forkball, fungo, knuckleball, shortstop*) Phrases: (e.g. *left-handed hitter, bat lag, arm side, pitch count, center field, breaking ball, baseball tee drill, Little League, Tommy John*)

138

Appendix 2

Technical Spoken Baseball Wordlist

1. arm
2. attaboy
3. away
4. backhand
5. backspin
6. bag
7. balance
8. balk
9. ball
10. ballpark
11. barrel
12. base
13. baseball
14. baseman
15. baserunner
16. bat
17. batter
18. bounce
19. BP
20. bucket
21. bullpen
22. bunt
23. cage
24. catcher
25. championship
26. changeup
27. clubbie
28. clubhouse
29. coach
30. consistency
31. contact
32. corner
33. count
34. curve
35. curveball
36. cutter
37. dealin
38. deceleration
39. defensively
40. delivery
41. depth
42. DH (designated hitter)
43. dinger
44. dirt
45. DL (disable list)
46. donut
47. double
48. drill
49. drive
50. dugout
51. extension
52. eye
53. fan
54. fastball

55. fence	94. out	133. slider
56. field	95. opposite	134. slump
57. fielder	96. outfield	135. slurve
58. fly	97. outfielder	136. speed
59. footwork	98. pick	137. spin
60. forkball	99. pitch	138. split
61. foul	100. pitcher	139. stadium
62. fundamentals	101. plate	140. stance
63. fungo	102. play	141. starter
64. game	103. player	142. stat
65. gamer	104. playoff	143. steal
66. glove	105. position	144. stealer
67. ground	106. postseason	145. stretch
68. groundball	107. practice	146. stride
69. hit	108. RBI (run batted in)	147. strike
70. hits	109. release	148. strikeout
71. hitter	110. replay	149. stuff
72. hole	111. rhythm	150. swing
73. hold	112. righty	151. switch
74. home	113. rotation	152. tag
75. hop	114. routine	153. target
76. hustle	115. rubber	154. tarp
77. infield	116. run	155. team
78. infielder	117. runner	156. teammate
79. inning	118. score	157. tee
80. inside	119. scoreboard	158. throw
81. knuckleball	120. scout	159. thrower
82. leadoff	121. screwball	160. tired
83. league	122. scrimmage	161. topspin
84. lefty	123. seam	162. toss
85. leg	124. seamer	163. ump
86. lineup	125. second	164. umpire
87. low	126. shag	165. velocity
88. miss	127. shift	166. walk
89. mound	128. short	167. wiffle
90. movement	129. shortstop	168. workout
91. mph	130. sign	169. zone
92. offseason	131. sinker	
93. oppo	132. slide	

139

Appendix 3

Top 100 (keyness value ranked) Technical Spoken Baseball Phrase list (for a complete list see Riccobono, 2018)

1.	second base	51.	breaking ball
2.	first base	52.	great baseball
3.	ground ball	53.	top outside corner
4.	home run	54.	young hitters
5.	strike zone	55.	bat lag
6.	line drive	56.	inside part
7.	home plate	57.	inner half
8.	big league	58.	little league
9.	foul ball	59.	stolen base
10.	weighted ball	60.	locker room
11.	baseball player	61.	right center
12.	long toss	62.	hand speed
13.	double play	63.	weight room
14.	power position	64.	pitch count
15.	third base	65.	nice stop
16.	inside move	66.	velocity killer
17.	release point	67.	palm ball
18.	right field	68.	bad throw
19.	center field	69.	split finger fastball
20.	next inning	70.	wristed pitches
21.	split finger	71.	arm side
22.	baseball tee	72.	neighborhood play
23.	shin guard	73.	baseball field
24.	fly ball	74.	position player
25.	base runner	75.	perfect game
26.	first pitch	76.	pro ball
27.	batting practice	77.	first inning
28.	pitching velocity	78.	first baseman
29.	nice pitch	79.	ball game
30.	swing path	80.	pitching drills
31.	arm slot	81.	jab step
32.	circle change	82.	lot of speed
33.	bat speed	83.	knuckle curve
34.	seam fastball	84.	baseball tee drill
35.	effective pitch	85.	bunt defense
36.	great play	86.	base stealer
37.	outside corner	87.	fair ball
38.	snow cone	88.	knuckle curveball
39.	short hop	89.	fastball arm
40.	second baseman	90.	particular pitch
41.	outer half	91.	handed hitter
42.	batting strength	92.	front toss
43.	home run derby	93.	bunt sign
44.	contact position	94.	arm action
45.	ball travel	95.	professional baseball player
46.	arm angle	96.	orange donut
47.	outside part	97.	bucket top
48.	right hander	98.	right-handed pitcher
49.	fence drill	99.	warning track
50.	glove side	100.	third baseman

140