High Frequency Words in Spoken English as a Lingua Franca in Academic Settings

Leah Gilner*

Abstract

This paper presents preliminary analyses of the coverage that the ICE-CORE word list provides of the Corpus of English as a Lingua Franca in Academic Settings (ELFA). The objective of this investigation was to assess the extent to which speakers' vocabulary preferences in this setting coincide with those of speakers' in other settings. Findings show that the dominant vocabulary in ELFA is the same as the one found in other corpora, whether in localized or globalized settings.

1. Background

This paper reports on a follow-up to preliminary analyses on the lexical distributions of vocabulary preferences of English speakers in localized and globalized settings. Specifically, the interest is centered on the use of high frequency words (HFWs) as they are represented in the ICE-CORE word list. The core analyses of this study center on the Corpus of English as a Lingua Franca in Academic Settings (ELFA). Additional supporting analyses have been conducted on the Vienna-Oxford International Corpus of English (VOICE), the International Corpus of English (ICE), and a collection of samples from 26 varieties of English. Each of these resources will be succinctly introduced hereafter.

The Corpus of English as a Lingua Franca in Academic Settings (ELFA) was created under the supervision of Prof Anna Mauranen at the University of Helsinki to meet the need for a means of investigating English as it is used by the international academic community to discuss, disseminate and exchange knowledge, findings and criticism worldwide. Briefly, ELFA contains transcriptions of approximately 131 hours of naturally occurring academic ELF in both monologic and dialogic speech events drawn from English-medium instruction graduate courses in Finnish universities as well as professional academic ELF as used in seminars, conferences, and doctoral defenses

* 准教授／Applied linguistics

(Carey, 2013). Speakers come from more than 50 countries and are described as well educated, plurilingual, and of variable proficiency.

ELFA amounts to slightly over 1 million transcribed words. It is structured according to disciplinary domain as well as according to speech event type. The domains are Social sciences (29%), Technology (19%), Humanities (17%), Natural sciences (13%), Medicine (10%), Behavioral sciences (7%), and Economics and Administration (5%). The speech event types are: lectures (14%), presentations (19%), and discussions (67%). Speech event types further subdivide according to setting (conference, doctoral defense, lecture, panel, and seminar) as well as clustering into monologic and polylogic events.

The Vienna-Oxford International Corpus of English (VOICE) was the first electronic compilation of data seeking to provide a representative samples of spoken English as a lingua franca. The corpus contains transcripts of naturally occurring, non-scripted face-to-face interactions. It amounts to about 1 million words of spoken ELF, corresponding to approximately 120 hours of transcribed speech. The speakers recorded for VOICE are experienced ELF users from a wide range of first language backgrounds. VOICE includes samples from 752 individuals, mainly from European countries, with approximately 50 different first languages (Corpus Description, 2013). The corpus data has been categorized into domains (professional, educational, leisure) as well as into speech event types (conversation, interview, meeting, panel, press conference, question-answer session, seminar discussion, service encounter, working group discussion, workshop discussion).

In the words of its managing director, “the International Corpus of English (ICE) began in 1990 with the primary aim of collecting material for comparative studies of English worldwide” (Nelson, 2011). To ensure coherence among individual corpora, ICE enforces certain guidelines. Specifically, each corpus contains 500 texts of approximately 2,000 words each collected from 1990 on; approximately 60% of a given corpus samples reflect spoken discourse represented by 100 private and 80 public dialogs as well as 120 scripted and unscripted monologues while approximately 40% of the samples capture written discourse in the form of 30 letters and 20 student writings along with 150 printed texts originating in instructional, academic, literary, newspaper, and other domains. Speakers are both male and female, 18 years old or older, and educated in the respective country. In this manner, ICE provides a means of analyzing spoken and written discourse of a particular variety (Nelson, 2011).

The collection of samples from 26 English varieties was the result of a centralized, coordinated, and relatively inexpensive effort by the author and colleagues (Gilner, Morales, and Shiobara, 2012). The collection amounts to 7,800 texts and approximately 15 million words organized in 26 sub-collections of 300 texts each. The compilation process followed the principles outlined by Sinclair (2004); in particular, every effort was made in order to represent a balanced view of each variety in

terms of mode, type, domain, language, location, and date. Furthermore, the contents were selected according to the communicative function in the community in which they arose rather than for the language they contained. The collection accounts for three types of discourse for each of the varieties under investigation, specifically, government documents, newspaper articles, and opinion columns. The English varieties sampled were from: Australia, the Bahamas, Belize, Bermuda, Cameroon, Canada, Fiji, India, Ireland, Jamaica, Kenya, Liberia, Malawi, Malaysia, Myanmar, New Zealand, Nigeria, Pakistan, the Philippines, Singapore, South Africa, Sri Lanka, Trinidad and Tobago, the UK, the USA, and Uganda.

The ICE-CORE word list originates from the analysis and comparison of the lexical distributions in seven corpora from the International Corpus of English (ICE), namely: Canada, East Africa, Hong Kong, India, Jamaica, the Philippines, and Singapore. Briefly, the methodology used to elicit the ICE-CORE was as follows. First, the lexical distributions for each corpus were calculated, producing a variety-specific frequency list for each. Second, each variety-specific list was used to assess its corresponding corpus. Third, through a number of iterations, the ICE-CORE word list came to contain a relatively sophisticated intersection of these variety-specific lists. In its original form, the ICE-CORE is composed of 1,206 word families.

It should be noted that, during the initial compilation of the ICE-CORE, a decision was made not to include cardinal and ordinal numbers, days of the week, and months of the year despite being both HFWs and common across the varieties investigated. The justification was based on the impossibility of grouping these words according to word families as these words are not related in a derivational or inflectional sense. Rather, these words can be grouped according to semantic categories and the inclusion of these categories alongside word families would open the door to other semantic categories such as, for instance, colors.

At the time, it seemed sound to avoid compromising the conceptual integrity of the list. Experience has shown that this approach introduces questionable gaps into the analyses. When profiling corpora, these four categories account for approximately from 1% to 3% of the running words. If not included, they need to be either added to a stop-list or marked as off-listed. The first case requires their removal from all counts while the second case requires these words to be considered unknown. Neither is satisfactory since neither provides an accurate reflection of language use. Yet, these issues notwithstanding, the word-family-only ICE-CORE word list has been used in all investigations to date (Gilner and Morales, 2011; Gilner et al., 2012; Gilner, 2014).

2. Methodology

The ELFA corpus is available in xml and text formats. The latter structures the content through the use of a set of xml-like custom tags. This format was deemed sufficiently informative

and was selected for the analyses hereby presented. Custom software was developed to clean, parse, and process the 165 files that make up the corpus into multiple overlapping clusters and categories. Additional profiling software was developed to elicit lexical distributions and other metrics.

As mentioned, the composition of the ICE-CORE has been found to be wanting. For this investigation, the ICE-CORE has been revised to include cardinal and ordinal numbers, days of the week, and months of the year. Since previous investigations used the unrevised ICE-CORE word list, all previous analyses on VOICE and other corpora have been carried out again for this paper. Thus, results below indicate a slighter higher coverage than those reported previously, a finding that coincides with expectations.

3. Results and discussion

The structure of ELFA makes it possible to divide the corpus into two umbrella categories of speech events: monologic and polylogic. This categorization is described as a language-internal one and refers to the number of speakers (i.e. single vs. multiple) involved in a given speech event (Mauranen, 2006). Table 1 shows the coverage the ICE-CORE provides of each of these event types as well as of ELFA as a whole.

	Coverage	Words	Events	Avg. length
Monologic	87.9%	332,401	91	3,653
Polylogic	91.38%	684,662	74	9,252
Entire corpus	90.24%	1,017,063	165	6,164

Table 1　HFW coverage of ELFA based speech event types

The first observation is that HFWs account for slightly over 9 out of 10 words of the entire corpus. That is to say, speakers of English as a lingua franca in academic settings demonstrate marked vocabulary preferences. Second, these preferences, in the form of a specific set of words, coincide with those found in previous investigations and involving speakers of English in localized and globalized settings (Gilner and Morales, 2011; Gilner et al., 2012). Third and last, results add support to the notion that these specific words are used even more frequently in globalized settings than they are in localized settings (Gilner, 2014).

Monologic speech event types are represented by lectures and presentations (see Table 2). These event types are mostly or uniquely lead by a single speaker, involving longer turns and control of the floor. The use of HFWs in monologic speech events, while still strikingly dominant, is 3.48% below that of polylogic events. Table 2 shows the breakdown of ELFA according to speech event types. Coverage data further specifies how discussions (polylogic event types) tend to make comparatively higher use of HFWs. As logically related pairs, conference presentations and

discussions are shown to rely the most on HFWs while discussions and presentations of doctoral defenses rely the least. The lowest coverage representative of speaker preferences across ELFA indicates a minimum of 8.5 words out of every 10.

	Coverage	Words	Type	Events	Avg. length
Conference discussion	92.03%	72,350	poly	14	5,168
Conference presentation	89.48%	93,267	mono	34	2,743
Doctoral defense discussion	90.62%	205,155	poly	14	14,654
Doctoral defense presentation	85.07%	21,743	mono	10	2,174
Lecture	87.08%	139,989	mono	20	6,999
Lecture discussion	91.02%	56,588	poly	12	4,716
Panel discussion	93.51%	13,064	poly	1	13,064
Seminar discussion	91.69%	337,505	poly	33	10,227
Seminar presentation	88.29%	77,402	mono	27	2,867
Entire corpus	90.24%	1,017,063		165	6,164

Table 2　HFW coverage of ELFA speech event types

Table 2 fleshes out the breakdown shown in Table 1. Although polylogic event types account for less than half of all events (44.85%), they are substantially longer on average, up to 6.74 times in the case of discussion versus presentation of doctoral defenses. Thus, the polylogic component of ELFA is dominant, showing a bias that is acknowledged to be deliberate. According to the corpus compilers, this decision is based on the insight that "it is in dialogic interaction that language primarily and most naturally gets negotiated" (Mauranen, 2006, p.153). Specifically, polylogic speech event types account for 67.31% of the running words in the ELFA corpus and are represented by discussions in seminars, conferences, panels, and doctoral defenses. On average, the use of HFWs in polylogic speech events is 1.14% over that of the entire corpus.

The comparatively higher presence of HFWs in polylogic speech events could be explained in at least two complementary ways. First, there could be a shift of accommodation strategies in monologic speech events when actual interlocutors are substituted by an imagined one, namely, oneself. The idea is that actual interlocutors offer linguistic material – in this case, words and phrases – that the speaker can reuse since they have been established to be known, thus leading to the recycling of certain vocabulary. In a monologue, and no matter how informative the non-linguistic feedback, the speaker is engaged with an imagined interlocutor (oneself), forcing the speaker to make guesses regarding which vocabulary is shared and known. Accommodation strategies in polylogic speech events are well documented and involve recycling, paraphrasing, repetition, and so on (Cogo, 2010; Cogo and Dewey, 2007; Jenkins, Cogo, and Dewey, 2011). The insight here is that these strategies can result in a narrower selection of vocabulary which, in turn,

may account for the conspicuous preference for HFWs shown in Tables 1 and 2. In monologic speech events, these guesses seem to result in a lesser reliance on HFWs.

Second, monologues might allow the speaker to rely on stretches of rehearsed speech, at least partially, while discussions have a real-time interactional component that forces interlocutors to improvise and produce spontaneous, unscripted discourse. The increased cognitive demands of this latter task compromise online resources and accessibility to less primed lexical items, thus producing fewer vocabulary flourishes. In contrast, prior preparation of material in academic settings is commonplace and one could argue that no worthy formal presentation (the prototypical monologic speech event in ELFA) is ever truly improvised. It can be posited then that premeditation and rehearsals result in a more diverse vocabulary that parallels the comparative increase in lexical diversity observed in written versus spoken discourse (Nation, 2006). If this were the case, the notion of planned versus unplanned discourse could be at play. Georgakopoulou and Goutsos (1997) explain that "….writers have time to mould their ideas into a more complex, coherent and integrated whole, making use of complicated lexical and syntactic devices" (p.34). Hatch (1992) had previously remarked that "…as writers revise and polish their performance, the language they use changes." (p.237) while, in spontaneous speech, "…words and phrases are repeated, and words seem to touch off the use of words having similar sound sequences" (p.241). Furthermore, the argument is that there is a bias towards redundancy in spoken discourse and that repeated words and phrases serve to promote cohesion. The lexical preferences exhibited in Tables 1 and 2 might be a manifestation of these communicative strategies.

The breakdown of ELFA according to domain was carried in terms of high level disciplines. According to the corpus compilers, the decision to use broad categories arose from the size of the samples and, in turn, of the corpus because, "…otherwise the search results remain too meagre" (Mauranen, 2006 p.152). Table 3 shows the extent to which HFWs account for the vocabulary in these domains.

	Coverage	Words	Events	Avg. length
Behavioral sciences	92.36%	72,377	10	7,238
Economics and Administration	90.65%	53,710	12	4,476
Humanities	91.23%	170,261	30	5,675
Medicine	85.68%	98,177	17	5,775
Natural sciences	88.81%	135,119	15	9,008
Other	93.51%	13,064	1	13,064
Social sciences	91.03%	279,960	50	5,599
Technology	90.44%	194,395	30	6,480
Entire corpus	90.24%	1,017,063	165	6,164

Table 3　HFW coverage of ELFA disciplinary domains

Research into why speakers within a particular domain rely more heavily on HFWs than speakers within other domains is beyond the scope of this investigation and, possibly, beyond the scope of ELFA. Nonetheless, it is worth noting that coverage numbers across domains reveal a surprising insight into vocabulary preference that is particularly remarkable considering how highly specialized these domains are: Approximately 1,200 word families dominate lexical distributions no matter the disciplinary field. This is noteworthy because we are instinctively aware of the fact that a doctoral defense in biology is fundamentally different than one in economics. It is unequivocally the case, it is in fact a tautology, that there is a plethora of linguistic elements that make it impossible to confuse one type of defense with the other and that among these linguistic elements is the choice of vocabulary, the terminology involved. Yet, findings show that more than 9 out 10 words are repetitions of members of the same 1,200 word families regardless of the domain. From a lexical perspective, speakers are able to express and elaborate with precision upon the radical conceptual differences between domains by means of less than 1 word out of every 10.

As will be shown shortly, these coverage numbers are somewhat higher than those found in localized, colingual, non-specialized speech. This could imply that the high degree of HFW usage is not idiosyncratic to the specialized discourse of these domains but, rather, a strategy employed by speakers of ELF. In other words, it could be inherent to the language users themselves rather than to the tasks in which they are engaged.

Before turning to data from localized settings, Table 4 lends support to this hypothesis by showing HFW coverage of VOICE alongside that of ELFA. As mentioned in the methodology section, VOICE was reanalyzed for this study with the modified ICE-CORE word list and this modification has yielded a higher presence of HFWs than that originally reported in Gilner (2014), specifically from 91.08% to 92.84%, an increase of 2.76%. Results provide at least weak confirmation of the significant preference for HFWs demonstrated by ELF speakers.

	Coverage	Words
ELFA	90.24%	1,017,063
VOICE	92.84%	981,515

Table 4　HFW coverage of two ELF corpora

The role of HFWs in both corpora is remarkable, coverage of VOICE perhaps reflecting the more general nature of the interactions sampled. Evidently, this is not sufficient to make strong claims regarding specific vocabulary preferences of ELF speakers in contrast to other populations in other settings. However, the tendencies are unmistakable. Those readers experienced with the lexical distributions of HFWs will immediately recognize that such high coverage numbers are unusual even if not entirely unheard of.

Table 5 shows vocabulary preferences in ICE by variety. Since both ELFA and VOICE are spoken-only corpora, data for the spoken component of each ICE corpus is also provided. The column labeled “Total” shows the size and HFW coverage of both the spoken and written component of each variety. The last row, labeled “Combined”, provides results for all 7 varieties combined.

	Spoken		Total	
	Coverage	Words	Coverage	Words
Canada	88.81%	642,280	85.49%	1,069,974
East Africa	87.26%	515,047	85.77%	1,407,342
Hong Kong	88.44%	969,707	86.22%	1,452,303
India	86.89%	685,376	84.02%	1,121,542
Jamaica	88.95%	654,581	86.00%	1,065,946
Philippines	86.40%	683,729	83.93%	1,128,509
Singapore	88.22%	665,021	85.73%	1,095,896
Combined	87.91%	4,815,741	85.35%	8,341,512

Table 5　HFW coverage of 7 varieties of English (localized use)

As with the VOICE data, ICE was reanalyzed for this study using the modified ICE-CORE word list. Coverage numbers are slightly higher than those reported in Gilner (2014).

The first observation is that spoken numbers are slightly higher than total numbers. This is uncontroversial. Analyses of corpora of written and spoken language consistently yield lower HFW usage for the first as the manufacture of written language allows for a more deliberate choice of vocabulary (Nation, 2006).

The second observation is that HFW coverage of ICE, whether as a whole or by variety, is lesser than that of ELFA and VOICE. The difference is sufficient enough to be statistically significant. Even considering the spoken component alone, the ICE average is 2.33% lower than ELFA and, and 4.93% lower than VOICE.

These numbers again lend support to the hypothesis that ELF speakers may have a comparatively higher preference for HFWs. However, it is important to note that the differences observed could be either introduced by the corpora themselves or be influenced by their design to an extent that shows marked biases. Were this to be the case, it would not be an argument against the representative adequacy of these corpora in general but, rather, a comment on the utility of these corpora for this specific type of analyses. As the work is preliminary, all options remain open.

For the sake of comprehensiveness, the 14 million-word-26-varieties collection was also reanalyzed with the modified ICE-CORE word list. Again, the HFW coverage originally reported in Gilner et al. (2012) is lower, the combined value found at 79.82%, than here at 82.74% (a difference of 2.92%).

	Coverage	Words		Coverage	Words
Australia	85.73%	556,040	Malaysia	79.32%	482,740
Bahamas	85.45%	553,281	Myanmar	78.80%	413,505
Belize	82.95%	503,681	New Zealand	86.95%	568,080
Bermuda	85.44%	551,795	Nigeria	83.58%	572,992
Cameroon	79.86%	355,386	Pakistan	79.09%	461,876
Canada	82.40%	614,284	Philippines	78.58%	576,299
Fiji	81.69%	523,760	Singapore	84.61%	588,605
India	79.52%	498,808	South Africa	81.88%	617,674
Ireland	84.85%	594,047	Sri Lanka	81.78%	550,973
Jamaica	81.92%	509,964	Trinidad and Tobago	85.59%	525,680
Kenya	84.31%	539,746	Uganda	79.66%	470,872
Liberia	81.07%	554,544	UK	85.17%	648,562
Malawi	82.36%	481,456	USA	83.97%	666,471
			Combined	82.74%	13,981,121

Table 6　HFW coverage of 26 varieties of English

This collection is noteworthy because its source is widely divergent from other corpora. Succinctly, each variety contains 300 documents from three equally represented domains. The government domain is composed of parliamentary Hansards or, when not available or sufficient in number, court rulings. The newspaper domain involves reporting articles that, while using less specialized speech, are nonetheless representative of a very specific genre. Last, the opinion domain again reflects a distinct type of discourse and, importantly, samples were gathered together with the entire, unedited comment threads. As a whole, the material in the collection ranges from highly formalized to highly volatile and, from a lexical point of view, rarified vocabulary is given as much importance as code-switching and fleeting internet coinages. On its own, it is remarkable that HFWs still dominate the vocabulary preferences of speakers no matter how diverse their geographic, political, and demographic locations and the situations in which they find themselves.

HFW coverage of this collection also shows the greatest disparity with ELFA and VOICE. The combined value, as shown in Table 6, is at 82.74% while those of ELFA and VOICE are at 90.24% and 92.84% respectively (corresponding to differences of 7.5% and 10.1%). In concrete terms, it means that when speakers find themselves in localized, colingual settings, 8 out of every 10 words are repetitions of approximately 1,200 word families. When these same speakers find themselves in globalized settings, this number increases to 9 out of every 10 words.

The simplest explanation is, again, the one provided by accommodation strategies. Much has been said about the individual uniqueness of English varieties as a reflection of the cultures from which they emerge. The difficulties inherent to communication when these barriers are erected

only serves to highlight the higher communicative value of what is shared over what might be more precise, elaborate, pertinent, or possibly territorial.

4. Conclusion

The findings from these analyses of ELFA support those obtained from the analyses of VOICE. When using English as lingua franca, speakers rely even more heavily on HFWs than they do when communicating within their local speech communities. This may be due to the "co-constructive, listener-oriented" nature of ELF interactions (Jenkins et al., 2011).

Any lingua franca serves the functional role of a shared vehicle of communication that goes beyond the boundaries of one community. Lingua franca (LF) interactions involve individuals from various and diverse linguistic and cultural backgrounds each of whom embody unique personal experiences. These same individuals, when interacting locally, draw on contextual, societal, and cultural conventions that may not be shared by their partners in LF interactions. Cogo (2010) describes LF exchanges as those "...where people from various backgrounds in more or less stable communities engage in communicative practices that shape, construct and define the communities themselves" (p. 296). Norms are negotiated by the participants for specific purposes by establishing a 'shared repertoire' of resources that promotes mutual understanding.

Findings presented here suggest that the HFWs are one of the features of the shared repertoire of resources when it comes to ELF. Carey (2013) discusses results of an analysis of frequent formulaic chunks that further evidences the preference of ELF speakers for HFWs. Carey identified the most frequently occurring three- to five-word chunks in the ELFA corpus and found, among other things, that speakers prefer to replace less frequent items with more frequent ones. For example, ELF speakers tend to use *so to say* rather than *so to speak*, replacing the less frequently-occurring *speak* with the more frequently-occurring *say*. Similar distributions of *so to say* and *so to speak* were observed in a supplementary analysis of VOICE.

Pitzl, Breiteneder, and Klimpfinger (2008) present findings into lexical innovations that illustrate another way that ELF speakers exploit HFWs to create "supportive and co-productive" (p. 40) interactional environments. Pitzl et al. focused on lexical variations found in a small subcorpus of VOICE as identified by the <pvc> tag. The VOICE corpus uses the <pvc> tag for individual lexical items that were not found in the reference dictionary. The researchers approached this preliminary analysis from the perspective of word formation and made observations based on the use of processes such as affixation, borrowing, analogy, and reanalysis. It was found that many of the lexical innovations in the subcorpus exploit HFWs as, for example, base forms for affixation. Examples include *increase*, *gather*, *imagine*, *prefer*, and *work*. The addition of a suffix often served to make meaning more overt and explicit. Mauranen (2007) and Ranta (2006) also observe that

ELF speakers tend toward communicative strategies (e.g. repetition, rephrasing, and discourse reflexivity) that make meaning more explicit. Other lexical innovations found in VOICE include uses of the prefixes *non-* and *re-* in combination with HFWs such as *formal*, *read*, *send*, and *confidence*. This tendency was attributed to the economy of expression whereby speakers minimize the number of words needed to express the idea they want to communicate. These findings provide further evidence of how ELF speakers accommodate each other by relying on a shared lexical resource while at the same time drawing on a sophisticated understanding of word formation potential.

The ELFA corpus is a welcome addition as well as a valuable contribution to corpora available to document English language use in the world today. This investigation adopted a feature-based descriptive approach with the aim of furthering our understanding of the role of HFWs in language use. The findings and discussion presented here indicate how this approach can complement the growing body of work focusing on the processes underlying interaction and meaning-making in ELF situations.

References

Carey, R. (2013). On the other side: formulaic organizing chunks in spoken and written academic ELF. Journal of English as a Lingua Franca, 2(2), 207–228. doi:10.1515/jelf-2013-0013

Cogo, A. (2010). Strategic use and Perceptions of English as a Lingua Franca. Poznan´ Studies in Contemporary Linguistics, 46(3), 295–312.

Cogo, A., & Dewey, M. (2006). Efficiency in ELF communication: From pragmatic motives to lexicogrammatical innovation. Nordic Journal of English Studies, 5(2), 59-93.

Corpus Description. (2013). The Vienna-Oxford International Corpus of English - Course Description. Retrieved from http://www.univie.ac.at/voice/page/corpus_description

Georgakopoulou, A., & Goutsos, D. (1997). Discourse analysis: An introduction. Edinburgh: Edinburgh University Press.

Gilner, L. (2014). An analysis of ELF speakers' lexical preferences. アジア英語研究 = Asian English Studies / 日本「アジア英語」学会 編, 16, 5–16.

Gilner, L. & Morales, F. (2011). The ICE-CORE word list: The lexical foundation of 7 varieties of English. Asian Englishes, 14(1), 4–21.Gilner, L., Morales, F., & Shiobara, K.. (2012). The creation of a corpus of 26 international varieties of English. 文京学院大学総合研究所紀要 / 文京学院大学総合研究所 編., (12), 35–43.

Hatch, E. M. (1992). Discourse and language education. Cambridge, England: Cambridge University Press.

Jenkins, J., Cogo, A., & Dewey, M. (2011). Review of developments in research into English as a lingua franca. Language Teaching, 44(3), 281–315.

Mauranen, A. (2006). A rich domain of ELF - The ELFA Corpus of Academic Discourse. Nordic Journal of English Studies, 5(2), 145–59.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? The Canadian Modern Language Review / La Revue Canadienne Des Langues Vivantes, 63, 1, 59-81.

Nelson, G. (2011). The International Corpus of English, from http://ice-corpora.net/ice/

Pitzl, M.L., Breiteneder, A. & Klimpfinger, T. (2008). A world of words: Processes of lexical innovation in VOICE. Views, 17(2), 21-46.

Ranta, Elina. (2006). The “attractive” progressive - Why use the -ing form in English as a lingua franca? Nordic Journal of English Studies, 5(2). Retrieved from http://hdl.handle.net/2077/3150

Sinclair, J. (2004). Corpus and text - Basic principles. In M. Wynne (Ed.), Developing linguistic corpora: A guide to good practice (pp. 1-16): Oxford: Oxbow Books. Available from http://ota.ahds.ac.uk/documents/creating/dlc/chapter1.htm [Accessed 2012-05-11].

（2014.9.26 受稿, 2014.11.19 受理）