Stress-testing the ICE-CORE Against 3.1 Billion Words

Leah Gilner*

Abstract

This paper presents preliminary analyses of the coverage that the ICE-CORE word list provides of a diverse range of English language texts. The objective of this investigation was to further test the limits of this word list. To date, the ICE-CORE word list has been used to profile corpora of spoken and written discourse of more than 25 international English varieties. Findings consistently indicate that very few words account for most of the language produced in local and global interactions. It is based on these and similar findings that it has been proposed that the ICE-CORE word list represents the lexical core of the English language. This investigation aimed at seeking the limits of this proposal by examining widely different corpora which contain texts prepared for very different purposes and audiences.

1.0 Introduction

The ICE-CORE word list is a modern embodiment of a well-established approach to vocabulary learning (Nation, 2001). It contains 1,206 words that are used extensively by English speakers around the world in both speech and writing. Briefly put, the items on this word list were originally identified by comparing corpora representing seven Inner and Outer circle English varieties, specifically, from Canada, East Africa, Hong Kong, India, Jamaica, the Philippines, and Singapore. The words that occurred with similar high frequency in all of these varieties were included on the ICE-CORE word list.

Since its inception, the ICE-CORE word list has been tested on a variety of corpora (Gilner, 2008; Gilner & Morales, 2011; Gilner, Morales, & Shiobara, 2012; Gilner, 2013). The initial work was done against the ICE corpus from which it originates. Subsequent analyses were conducted against the 26 English varieties collection (Gilner et al., 2012) and the VOICE corpus (Gilner, 2013). All work to date has shown that the ICE-CORE wordlist accounts for 75%-90% of the lexical choices made by worldwide English speakers in colingual as well as international settings

＊准教授／応用言語学

(Gilner & Morales, 2011; Gilner, Morales, & Shiobara, 2012; Gilner, 2013). Therefore, it is so far safe to conclude that the ICE-CORE is representative of English language use and, consequently, of significant pedagogical value.

Despite the breadth of domains and registers analyzed, there remains much to be tested. To this end, the following three lines of inquiry have been opened:

It is of interest to determine the function the ICE-CORE wordlist plays in graded readers as these are, or should be, among the first models learners are exposed to from early on in their language learning endeavor.

As learners' proficiency improves, they should be exposed to models that are not explicitly designed to serve their specific needs but rather to distribute information to wider, layperson audiences. Nowadays, the internet makes available to the general population a great deal of such resources. Due to its open nature, Wikimedia is an excellent starting point for this kind of analysis.

A language learning endeavor could and possibly should lead to incorporation of the acquired language into the personal and the professional life of the learner. It is of interest, therefore, to measure the extent to which the ICE-CORE list is adequate in the pursuit of higher academic skills and knowledge as well as in professional formation.

In this paper, a brief summary of preliminary results on these three avenues of research is provided together with a number of considerations regarding the nature and limitations of the available data. Some observations on the working hypotheses will be elaborated upon whenever appropriate. The objective of the paper is therefore best described as an early report for colleagues with an interest in the subject as well as a primitive roadmap for future research.

2.0 Methodology

The ICE-CORE word list was compiled by means of corpus analysis. The words on the list were identified through the analysis and comparison of the lexical distributions in seven corpora compiled by the International Corpus of English (ICE). "The International Corpus of English began in 1990 with the primary aim of collecting material for comparative studies of English worldwide" (Nelson, G., 2010). The varieties analyzed were: Canada (Nelson, J., 2010), East Africa (Hudson-Ettle, and Schmied, 2005), Hong Kong (University College and The Chinese University of Hong Kong, 2006), India (Shastri and Leitner, 2002), Jamaica (Rosenfelder, Jantos, Höhn, & Mair, 2009), Philippines (University College and De La Salle University, 2002), and Singapore (University College and National University of Singapore, 2005).

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; 70% of a given corpus reflects spoken discourse represented by 100 private and 80 public dialogs as well

as 120 unscripted monologues while 30% captures 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 through the medium of English in the respective country. In this manner, ICE provides a means of analyzing spoken and written discourse of a particular variety (Nelson, G., 2010).

For this paper, the ICE-CORE word list was used to profile three diverse collections of texts. First, the Grader Reader Corpus totals approximately 1.5 million running words. It represents a widely used resource in the domain of English education. Second, the Wikimedia Corpus totals approximately 1.6 billion running words. It represents easily-accessible online materials that have become frequently used as reference sources. Third, the Academic and Professional Literature Corpus totals approximately 1.5 billion running words. It contains didactic reading material that is used by students and professionals who need to further their understanding in a particular field.

It should be noted that the term 'corpus' has been used liberally throughout this paper to refer to these collections of texts. It must be kept in mind that the compilation process used to gather the samples for each of these collections lacks the rigor and balance required to qualify them as corpora in the technical sense. Rather, these collections were compiled opportunistically as texts crossed the path of the author over many years. The usefulness of these collections, therefore, is narrowed down to finding the limitations of the ICE-CORE as well as a means to start developing a degree of familiarity with the domains from which these texts come.

Each of the analyses reported below followed the same procedure. The integrity of the electronic data contained in each collection was confirmed through cleaning and parsing procedures implemented with custom software. At this point, profiling software (also custom-made for the job) was used to profile each corpus against the ICE-CORE word list.

3.0 Analysis and results

3.1 The Graded Reader Corpus

The first line of inquiry requires the analysis of graded readers. To this end, a collection of 126 texts from 4 publishers has been compiled. These add up to approximately 1.5 million words. The graded reader corpus is relatively small by today's standards. Corpora into the tens of millions of words have become common place and this very paper reports below on corpora that reach the billion and a half word mark. However, it is relevant to point out that the relatively small size of this corpus is nonetheless not a make-or-break characteristic. Useful insights can and have been obtained on corpora of this size and even smaller (e.g. the Brown Family Corpora, the International Corpus of English, etc).

Table 1 presents a basic skeleton of the characteristics of the graded reader corpus together

with the results of the first ICE-CORE coverage analysis.

	Texts	Words	Average length	Coverage
Publisher A	80	1,172,431	14,655	86.85%
Publisher B	8	71,677	8,960	89.61%
Publisher C	19	192,418	10,127	85.28%
Publisher D	9	131,562	14,618	82.45%
All	116	1,568,088	13,518	86.44%

Table 1 The Graded Reader Corpus and ICE-CORE coverage.

As shown, the collection is uneven. There are, for example, as many as 10 times more texts from Publisher A than from Publisher B. The average length of these readers also varies greatly from publisher to publisher. Moreover, there is only one complete collection of readers with all levels included as the publisher intended (Publisher A). These shortcomings are difficult to compensate for due to the fact that publishers are generally unwilling to make their materials available for analysis. It is therefore important to state so clearly from the start.

Regarding token coverage by the ICE-CORE wordlist, those familiar with existing research will not be surprised at all to see percentage values into the mid 80's. This wordlist has been shown to be robust in previous analyses and there is no reason to think it would fail here. In essence, the coverage numbers show that 8 to 9 words from every 10 belong to the ICE-CORE and, consequently, would simultaneously/reciprocally be learned by a student through exposure to the readers and the wordlist. One activity would support the other, in other words.

As mentioned, Publisher A is represented in this corpus by an entire collection. This makes it possible to analyze the texts according to level. Table 2 presents a summary of results for this intriguing proposition.

	Texts	Words	Average length	Coverage
Level 1	10	19,361	1,936	83.95%
Level 2	17	104,070	6,122	85.29%
Level 3	17	169,693	9,982	85.66%
Level 4	16	305,361	19,085	87.85%
Level 5	20	573,946	28,697	87.03%
All	80	1,172,431	14,655	86.85%

Table 2 A graded reader collection by levels.

The publisher chose to include fewer texts in the first level and more in the last. As it is often the case, there is no information regarding this decision and its objectives. There is however expected

uniformity in the progressive increment of average text length as the level of difficulty increases. This is normal and welcome. As learners progress, they can handle longer, more complex passages and readings. The coverage results, on the other hand, are somewhat counterintuitive. Although overall significant, ICE-CORE coverage shows an inverse relation to level of difficulty. At this point, and since the publisher does not provide a defining vocabulary or rationale for its decisions in this respect, the issue can only be addressed through the formulation of likely scenarios.

The most likely scenario is that the defining vocabulary used in the first level is a subset of the ICE-CORE and, as difficulty increases, more and more ICE-CORE words are included. Another, although less likely explanation, is that the defining vocabulary is at odds with corpora information for reasons unknown and that, as levels increase, the relaxation of lexical criteria makes the readers more akin to standard language use and, consequently, a greater number of ICE-CORE words appear. A third possibility is the use in lower levels of greater number of less frequent words (thus, outside the ICE-CORE) due to their general less unambiguous semantic and collocational characteristics.

Summing up, analyses of graded readers are intrinsically difficult due to the roadblocks created by publishers. Subsequent analyses should try to determine the defining vocabulary by raw frequency parsing and range statistics. If the readers are properly designed, reoccurrence should be a fundamental characteristic of the presentation of the vocabulary and a key to the elicitation of the lexical learning strategy adopted by the publisher. Once this information has been elicited, and to the extent it is deemed to be reliable, it will be possible to establish the accuracy of the scenarios discussed previously. These insights would be of particular interest in order to determine the value of the ICE-CORE in the early stages of acquisition when graded readers have the most value.

3.2 The Wikimedia Corpus

The second line of inquiry addresses access to general information not specifically designed with language learners in mind. The reach, breadth, policies, and politics of Wikimedia make this informational resource particularly attractive. It should be said that Wikimedia and its best known appendage Wikipedia are not without reason neck deep in controversy due to a myriad of issues that would be beyond the scope of this paper to even begin to list. For the purpose of these analyses, however, these factors are of limited consequence because the ease of access via the internet together with its, perhaps undeserved, popularity makes the availability of Wikimedia information a real and present research interest. In sum, and despite all concerns, combined Wikimedia projects were the 9th most visited in the world in 2007, claiming more than 90 billion page views that year (Wikimedia Meta-Wiki, 2013), and allowing the leaders that run the venture to aim at 1 billion visitors per month by the year 2015 (Ting Chen, 2011). In other words, our students will inevitably

find their way to Wikimedia.

That said, the nature of the project makes access to the raw data possible but tortuous. Database dumps are made publically available and those from the first week of September 2013 were obtained for the Wikipedia, Wikinews, Wikivoyage, Wikibooks, Simplewiki, and Wikiversity projects. Table 3 shows a summary of the word counts for each together with ICE-CORE coverage statistics.

	Words	Coverage
Wikipedia	1,559,109,561	72.80%
Wikinews	5,711,832	76.12%
Wikivoyage	13,960,989	75.30%
Wikibooks	32,201,226	77.31%
Simplewiki	14,852,140	73.23%
Wikiversity	10,750,971	76.76%
All	1,636,586,719	75.25%

Table 3　The Wikimedia Corpus word counts and ICE-CORE coverage.

As mentioned, Wikimedia allows open access to their data. Unfortunately, this is a political policy with more interest in publicity than substance. The result is that identifying the actual content and extracting it from the dumps is a nightmare that can turn into an impossible mission at every turn. Moreover, Wikimedia's never-ending campaign of self-promotion makes their statistical metadata evidently suspect. For this reason, and until proper independent analyses can be conducted, information such as number and average length of articles has been removed from the table.

A number of the most significant disclaimers out of the way, the ICE-CORE coverage of Wikimedia projects is relatively low at 7 to 8 words for every 10. This, of course, is not an outright demerit because the language policies of Wikimedia are a mess (by necessity and design), making any form of lexical uniformity closer to a linguistic bazaar than to a formal publication. This situation amounts to a true stress-testing ground where less robust wordlists would be found wanting. The ICE-CORE survives since it provides significant coverage that would allow a learner working access to the information.

Summing up, Wikimedia presents a massive challenge for researchers. As with graded readers, the creators of the material are unwilling to disclose the true intent and extent of success of execution. This again forces the sort of inquiry proposed here to start by first making analyses of the raw data before proceeding with useful coverage statistics. Nonetheless, Wikimedia is one of the giant internet reference tools if not the dominant provider of pseudo-formal information on nearly everything. Adequate understanding and characterization of this resource would be a

valuable contribution to language education.

3.3 The Academic and Professional Literature Corpus

The third line of inquiry addresses access to the most complex of all informational resources: academic and professional literature. Learners are bound to encounter these materials if their linguistic progress is non-trivial. Furthermore, it is important to note that this kind of materials present the greatest linguistic challenge for their readership, be those first or subsequent language users. A proper assessment of the lexical demands of formal language is therefore of great interest.

As shown in Table 4, nearly 10,000 texts from 16 disciplines have been collected for this preliminary analysis. They fall into three general categories across disciplines: college textbooks, reference volumes, and professional/technical documentation. Together, the 16 corpora amount to approximately 1.5 billion words. The largest corpus is Computer Science at almost 240 million words and 1,800 texts, followed by Biology, Mathematics, Physics, and Psychology. Together, these 5 corpora account for ~52% of the entire collection with Computer Science alone accounting for ~16% of the total. The smallest corpus is Economics with some 41 million words and 387 texts, closely followed by Electronics, Linguistics, Engineering, and Law. Together, these 5 corpora account for ~17% of the entire collection.

There is an average of 625 texts per discipline with a maximum of 1,796 Computer Science texts and a minimum of 191 Linguistics texts. The average length of a text is slightly above 170,000 words with a maximum at 281,423 words (Linguistics) and a minimum at 93,657 words (Mathematics).

Corpus	Texts	Words	Average length	Coverage
Biology	765	166,683,115	217,886	65.42%
Chemistry	354	61,472,766	173,652	65.13%
Computer Science	1,796	238,913,897	133,026	76.56%
Economics	387	41,277,316	106,660	81.10%
Electronics	276	46,983,349	170,230	71.72%
Engineering	446	55,388,843	124,190	70.31%
History	463	87,731,308	189,484	72.80%
Law	378	59,040,433	156,192	77.97%
Linguistics	191	53,751,701	281,423	73.88%
Mathematics	1,609	150,694,049	93,657	71.19%
Medicine	364	81,801,936	224,731	66.37%
Nano Science	390	77,761,979	199,390	64.03%
Philosophy	586	74,876,765	127,776	79.79%
Physics	764	121,849,057	159,488	72.13%
Political Science	408	94,409,260	231,395	74.83%
Psychology	815	117,228,084	143,838	76.52%
All	9,992	1,529,863,858	170,814	72.48%

Table 4　Corpora from 16 academic/professional disciplines.

As shown, the collection is unbalanced across disciplines in number and length of texts. Not shown, the collection is also unbalanced internally as the number of college textbooks, reference volumes, and professional/technical documentation is different for each discipline. These problems aside for the time being, and notwithstanding the additional strain that they impose on the profiling word list, the ICE-CORE performs manifestly well at somewhere above 7 out of every 10 words. For reasons to be discerned in future research, Economic texts are better served by the ICE-CORE at 81.10% while, perhaps unsurprisingly considering the highly technical degree implicit, Nano Science is worse represented at 64.03%.

Importantly, the average coverage is only a few percentage points below Wikimedia's. This is significant because this collection is composed of true formal materials that use language as methodically and unequivocally as possible in order to provide the most efficient explanatory and illustrative material. In this manner, the language of a psychology textbook could be considered obtuse by colloquial standards but, if it is properly written, it uses language in the most clear, concise, unambiguous way available to the authors and, to the extent that they succeeded in providing satisfactory explanations of psychological concepts, they do so by also unintentionally providing instruction of the very language used in those explanations.

This consideration lends itself to the weak formulation of a learnability hypothesis, namely, that linguistic acquisition is a necessary component of concept acquisition when this is mediated by language. The implication is that a word list extracted from these corpora could provide insights into lexical learnability, an otherwise murky concept that is perhaps one of the most intriguing questions still open in vocabulary acquisition.

Summing up, preliminary analyses of formal materials in higher education and professional development show that the ICE-CORE maintains its usefulness to learners. The intrinsic difficulty of these texts due to their specialized nature is apparently no obstacle to the notion that language users insist on doing a lot with little, making use of a relatively small set of words even when explaining the most elaborate and complex formal notions.

4.0 Closing remarks

As is often the case with interesting research, more questions have been generated than answers found. Additionally, a number of the limitations of these collections have been identified. These observations have been elaborated on as they arose throughout the discussion of results.

In closing, the ICE-CORE word list has been shown to be a valuable lexical resource for those interested in using graded readers, Wikimedia materials, and academic/professional literature. Mastery of the 1,206 words in the ICE-CORE will ensure the understanding of approximately seven or eight of every ten words in any of these texts. As has been noted repeatedly for over one

hundred years, usage analyses of the most frequent words in the language consistently yield a formidable display of versatility and usefulness.

References

Gilner, L. (2008). *The lexical foundation of English varieties*. Paper presented at the Japan Association for Asian Englishes 24th National Conference, Tokyo, Japan.

Gilner, L. (2013). *Identification of the lexicon preferred by speakers of English as a lingua franca*. Paper presented at the Japan Association for Asian Englishes 32nd National Conference, Osaka, Japan.

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). High frequency words in world Englishes. 文京学院大学総合研究所紀要第12号, pp. 99-111 [Bunkyo Gakuin University General Research Institute Journal 12]

Hudson-Ettle, D. M. & Schmied, J. (2005). *The International Corpus of English: The East Africa Component*. London: University College, Survey of English Usage, International Corpus of English.

Nelson, G. (2010). *International Corpus of English*. Retrieved from: http://ice-corpora.net/ice/

Nelson, J. (2010). *International Corpus of English: The Canadian component*. Edmonton: University of Alberta.

Rosenfelder, I., Jantos, S., Höhn, N., & Mair, C. (2009). *International Corpus of English: The Jamaican component*. Germany: University of Freiburg.

University College and The Chinese University of Hong Kong. (2006). International *Corpus of English: The Hong Kong Component*. London: University College, London Survey of English Usage.

University College and The National University of Singapore. (2005). *The International Corpus of English: The Singapore Component*. London: Survey of English Usage.

University College and De La Salle University. (2002) *The International Corpus of English: The Philippines Component*. London: Survey of English Usage.

Shastri, S. V. & Leitner, G. (2002). *The International Corpus of English: Indian Component*. London: University College, London Survey of English Usage.

Ting Chen. (2011). Wikimedia presents its five-year strategic plan. *Wikimedia Foundation*. Retrieved from http://blog.wikimedia.org/2011/02/25/wikimedia-presents-its-five-year-strategic-plan/

Wikimedia Meta-Wiki.(2013) Wikimedia in figures - Wikipedia. Retrieved from http://meta.wikimedia.org/wiki/Wikimedia_in_figures_-_Wikipedia

（2013.9.25 受稿 , 2013.11.27 受理）