Characteristics of UI English: From Non-native's Viewpoint

Ryutaro Nishino and Kayoko Nohara

Tokyo Institute of Technology, Tokyo, Japan
ryutaro@nishinos.com

Abstract. Multicultural aspects of user interfaces (UIs) have been studied for years. However, the characteristics of UI English from the viewpoint of non-native speakers of English have not been discussed widely. This study compares a UI English corpus with general English corpora using the word list coverage method and the lexical diversity method. It finds that UI English contains more words that are above introductory level but that it is less diverse than general English. Therefore, for non-natives to use software smoothly, they need to learn a relatively limited number of words that frequently appear in UIs.

Keywords: English, non-native, user interface, UI.

1 Introduction

Many software products have been developed by US-based companies and their user interfaces (UIs) are usually written in English. The Business Software Alliance says, "An estimated 65% of the PC software units in service worldwide in 2008 were from US-based companies." [1] More recently, distributing software globally through the Internet, such as web or smartphone applications, is becoming easier and easier.

English UIs are often translated and localized for global use. The importance of localization has been pointed out for years to market software products globally [2, 3]. Although some software applications are translated into users' languages, others are not due to cost or time constraints. As a result, many non-natives have a chance to use UIs written in English.

The first objective of this study is to clarify the characteristics of English currently used in UIs, from the viewpoint of how difficult the English is for non-natives. We see the difficulty from two points in this paper: 1) the coverage of a text by English word lists for learners and 2) the lexical or vocabulary diversity in a text. Clarifying the characteristics of UI English will eventually lead to better UIs, especially for non-natives. The second objective is to find characteristic words used in UIs. By learning these words, non-natives will be able to use software more comfortably. The second objective is rather a preliminary study for future work.

P.L.P. Rau (Ed.): CCD/HCII 2013, Part I, LNCS 8023, pp. 323–332, 2013.

2 Related Research

2.1 Studies of UI from Multicultural and Non-natives Perspectives

UIs have been studied from the multicultural perspective for years by different researchers [4-6]. However, UIs have not been discussed widely from the perspective of non-native speakers of English, while some researchers pointed out the importance in remarks such as "The user interface must be designed and written so that it is understandable by a user who speaks English as a second language." [7]

Because large software companies now recognize the importance of writing English that is understandable for non-native users [8-10], scientific studies need to be done to better understand the characteristics of UI English from the viewpoint of non-native speakers of English, in addition to the multicultural perspective.

2.2 English Word Lists for Learners

One of the methods to describe the characteristics of the vocabulary in an English text is to compare it with word lists used by learners of English. By measuring how many percentage of a vocabulary is covered by word lists, one can guess how difficult the text is for learners. Note, however, that this method checks the vocabulary knowledge only, and other kinds of abilities such as listening or speaking will not evaluated.

Examples of the frequently used word lists are General Service List (GSL) [11] and Academic Word List (AWL) [12]. The GSL was developed from a corpus largely based on the basis of frequency [13]. It includes a total of 2,000 word families[1] with two different levels, 1,000 for each. The AWL includes 570 word families that are not in the GSL and that are frequently used in academic texts. The GSL and AWL have been used in combination to see the coverage of text in fields such as business [14] or public health [15]. Another example of word lists is the list developed by Paul Nation from the British National Corpus (BNC) [16]. This BNC list is in the order of frequency and has 14 levels, 1,000 word families for each.

Although not being a word list like ones explained above, the English Vocabulary Profile (EVP) is a vocabulary database with six proficiency levels based on the Common European Framework of Reference [17]. The levels are Basic User (A1 and A2), Independent User (B1 and B2), and Proficient User (C1 and C2).

2.3 Lexical Diversity

To find the characteristics of a text, a measure called lexical diversity is often used. Lexical diversity shows how many different words are used in an entire text.

[1] A word family is a set of related words. For example, the headword "accept" has "acceptability," "acceptable," "unacceptable," "acceptance," "accepted," "accepting," and "accepts" as family members.

The oldest and most frequently used measure is type-token[2] ratio (TTR) [18]. The higher the TTR is, the more different words are used in a text. The problem in TTR is the dependence on text length [19]. As a text becomes longer, its TTR decreases because the same types are used repeatedly. To overcome this problem, other methods, such as Guiraud's index [20] or Herdan's index [21], have been proposed. However, such measures in practice depend on their text length [19], too. As a result, no index can reliably measure the lexical diversity between texts with different sizes. As a practical solution, Baayen [22] proposes to compare the number of types across texts with the same text sizes. In a recent study, different kinds of measures (TTR, Guiraud, Herdan, Uber, Maas, vocd-D, HD-D) are used in combination [18].

2.4 Keyword Extraction

There are several methods to extract terms specially used in a certain corpus. In recent years, comparing the numbers of appearance between a target corpus and a reference corpus is often used. In comparing corpora, different statistics are used to determine whether a certain word is specially used in a certain field.

There have been discussions of which statistic is valid or effective [23, 24]. Chujo and Utiyama [25] compare nine statistics for extracting keywords in the applied science field and claim that different statistics are effective to extract terms for learners with different proficiency levels. For example, the log-likelihood ratio (LLR) is effective for intermediate-level.

3 Methods

3.1 Preparing Corpora

Target Corpus. The target corpus is compiled from English used in UIs such as buttons or messages. We collected text from different genres of software that are well known and have many users so that the corpus represents the English actually seen and read by users. The genres of software are mobile, productivity, communication, and entertainment. We do not adjust the text size of each genre. Table 2 details the genre, name, and size. The text of Android applications is from Android Git repository [26] and the other text is from the Microsoft Language Portal [27].

Software text includes characters that should be removed or modified, such as HTML tags or variable placeholders. Because the types of such characters differ from software to software, we cleaned the files according to the basic policy: "Retain characters that are actually shown to users." Table 1 shows examples of what we did for the cleanup.

[2] Tokens are the total number of words in a text. Types are the number of unique words. In the sentence "I ate an apple that I bought yesterday", for instance, tokens are eight and types are seven (because "I" appears twice). The TTR in this case is 7/8 = 0.875 (87.5%).

Table 1. Modification of special characters in software

Type of character	*Example*	*Modification*
Variable placeholder	%d, %s1, {0}	Replaced with <variablehere> to indicate that something is displayed at runtime.
Metacharacter	/u2026, ¥n	Removed, or replaced with the character it represents.
HTML tag	, 	Removed, or replaced with the character it represents.
Keyboard shortcut	Hi&de	Removed unneeded character ("&" is removed in the example).
Words without space	%sAuthor	Added a space to make them an individual word ("%s Author" in the example).

We separately counted the tags that represent variable placeholders (i.e., <variablehere>) into the total number of words because we do not know if the word or sentence put here is known or unknown to the user until the software is actually executed. The final numbers of words after removal or replacement are in Table 2.

Table 2. Software in the UI corpus

Genre	*Application name*	*Size (number of words)*
Mobile	Android applications[3]	23,552
Productivity	Microsoft Word 2007 SP2	80,304
Productivity	Microsoft Excel 2007 SP2	109,821
Communication	Internet Explorer 8	147,814
Communication	Hotmail	27,266
Entertainment	Windows Media Player 10	18,762
		407,519

Reference Corpus. We used the Brown Corpus, the Manually Annotated Sub-Corpus (MASC) of Open American National Corpus (OANC), and the CNN Tech news corpus as reference corpora. The Brown Corpus is an American English corpus with about 1 million words from the prose printed in the U.S. during the year 1961 [28]. The MASC is a balanced subset of 500,000 words of written and spoken text from OANC [29]. The CNN Tech news corpus was compiled by the authors from the CNN news articles published in 2011 and categorized under "Tech" [30]. The CNN Tech

[3] Android applications: AccountsAndSyncSettings, AlarmClock, BasicSmsReceiver, Bluetooth, Browser, Calculator, Calendar, Camera, CellBroadcastReceiver, CertInstaller, Contacts, DeskClock, Email, Exchange, Gallery, Gallery2, Gallery3D, GlobalSearch, GoogleSearch, HTMLViewer, IM, KeyChain, Launcher, Launcher2, LegacyCamera, Mms, Music, MusicFX, Nfc, PackageInstaller, Phone, Protips, QuickSearchBox, Settings, SoundRecorder, SpareParts, SpeechRecorder, Stk, Tag, VideoEditor, VoiceDialer.

corpus has over 600,000 words, and many of the articles are IT-related topics including video games or consumer electronics, but excluding energy, environment, or space technologies. The reason why we use this corpus is to see whether or not IT-related terms, which are very often found in the UI corpus, affect the characteristics of the text. Thus, the first two corpora, Brown and MASC, are general American English corpus, and the last one, CNN Tech, is a corpus of a specific field.

3.2 Determining the Difficulty by Two Criteria

As stated in the introduction section, we see the difficulty for non-natives from two points: 1) the coverage by word lists and 2) the lexical diversity.

Coverage by Word Lists. We adopted the GSL+AWL and the BNC list by Nation to measure how much of a text is covered by the lists to estimate the vocabulary level. The GSL has two levels (Level 1 and 2) and the AWL is the extension of GSL (i.e., Level 3). The BNC list has 14 levels. We used these lists to check how much percentage of tokens in a corpus is covered by the lists. One can guess that the lower the coverage is, the more difficult the text is for non-natives because they need remember more words. We applied this method to different kinds of corpora to compare the relative levels of texts. The important assumption here is that non-natives learn vocabulary in the order of frequency (frequently-used words are learned first). To calculate the coverage, we used a software tool called AntWordProfiler [31].

Lexical Diversity. Because there is not a measure with established reputation, we here use the Guiraud's index (R) and the Herdan's index (C):

$$R = \frac{Types}{\sqrt{Tokens}} \tag{1}$$

$$C = \frac{\log_e Types}{\log_e Tokens} \tag{2}$$

We need to be careful that these indices depend on the text length (i.e., tokens). In addition to this, we compare texts after adjusting their sizes, as Baayen suggested [22]. We adjust the text sizes to the shortest one. The approximate sizes of the corpora we use are: the UI corpus: 400k, Brown: 1million, MASC: 500k, and CNN Tech: 630k. So, we cut the sizes to that of the UI corpus. First, we divide all corpora into the groups of 50k words. For instance, the UI corpus is divided into 8 groups (400k / 50k = 8). The reminder is simply discarded. Likewise, Brown is divided into 20 groups, MASC 10 groups, and CNN Tech 12 groups. And then, we randomly choose 8 groups from each of them to make corpora with 400k tokens.

We apply three methods (Guiraud, Herdan, and same-size comparison) to four corpora to compare their lexical diversity. The R software [32] and its package languageR [33] are used to count types and tokens. Before counting, we erased punctuation marks in the corpora. This is to avoid a word followed by such marks (e.g., "corpora.") to be recognized as one word. The following marks are replaced with a space:

```
.   ,   "   '   ?   !   (   )   [   ]   --
```

3.3 Finding Characteristic Words

We use the log-likelihood ratio (LLR) statistic to identify characteristic words used in the UI corpus. For a reference corpus, we combine the Brown and MASC corpora, which are general American English corpora, to make them an approximately 1.5 million-word corpus. The AntConc version 3.3.5 [34] is used to calculate LLR. Default settings are used, but a lemma list created by Yasumasa Someya (available from the AntConc website) is used with some modifications by the authors such as adding new words (e.g., app, email, etc.). No part-of-speech (POS) information is used.

After extracting words based on the LLR statistic, we manually pick up action-related words (mainly verbs) and create a list of top 10 words (due to the limitation of pages of this paper). The reason why we focus on action-related words here is that they describe human-computer interactions and are valuable compared to nouns that include many technology terms, which would disappear when new technologies appear. And then, we identify the proficiency levels for each word by using the GSL/AWL, the BNC list, and the English Vocabulary Profile (EVP).

4 Results

4.1 Coverage by Word Lists

Table 3 shows the coverage of the corpora by the GSL 1, GSL 2, AWL, variables (such as "%d" or "{0}"), and their total in percentage.

Table 3. Coverage by GSL+AWL (%)

	GSL 1	*GSL 2*	*AWL*	*Variables*	*Total*
UI corpus	**61.1**	6.5	**13.6**	1.4	82.7
Brown	76.5	5.7	4.7	-	86.9
MASC	74.4	5.4	4.8	-	84.6
CNN Tech	72.5	5.5	6.0	-	84.0

Table 4 shows the coverage of the corpora by the different levels of the BNC list, variables, and their total in percentage. Each level has 1,000 word families.

Table 4. Coverage by BNC list (%)

	BNC 1	*BNC 2*	*BNC 3*	*BNC 4*	*BNC 5*	*BNC 6-14*	*Variables*	*Total*
UI corpus	**67.5**	**12.0**	**5.0**	**3.8**	1.0	3.9	1.4	94.6
Brown	77.0	8.3	3.2	2.4	1.4	3.5	-	95.7
MASC	76.0	7.7	2.6	2.3	1.2	3.4	-	93.3
CNN Tech	75.3	8.6	2.8	2.6	0.9	3.2	-	93.3

4.2 Lexical Diversity

Table 5 shows the lexical diversity of the corpora using Guiraud and Herdan indices.

Table 5. Lexical diversity using Guiraud and Herdan indices

	Types	*Tokens*	*Guiraud*	*Herdan*
UI corpus	11,651	409,826[4]	**18.20**	**0.72**
Brown	47,059	1,025,536	46.47	0.78
MASC	35,537	500,227	50.25	0.80
CNN Tech	28,904	635,315	36.26	0.77

Table 6 shows the lexical diversity of the corpora with the same text size.

Table 6. Lexical diversity using fixed text sizes

	Types	*Tokens*	*Type-token ratio (%)*
UI corpus	11,395	400,000	**2.85**
Brown	28,737	400,000	7.18
MASC	30,916	400,000	7.73
CNN Tech	22,932	400,000	5.73

4.3 Characteristic Words

Table 7 lists the top 10 action-related words that are characteristic in the UI corpus compared with the general English corpus (Brown+MASC). The words are listed alphabetically with the proficiency levels on the word lists, and the LLR statistic.

Table 7. Top 10 characteristic action-related words

	GSL/AWL	*BNC*	*EVP*	*LLR*
add	1	1	A2	3409.8
allow	1	1	B1 or C1	3059.1
change	1	1	A2	3291.6
click	none	3	A2	4939.9
configure	none	8	none	3180.9
disable	none	none	none	4695.2
enable	3	2	B2	4263.2
select	3	2	B1	5072.1
set	1	1	B1 or B2	7339.2
use	1	1	A1	4508.1

[4] The total number is slightly different from the one in Table 2 because different tools are used. Here we used languageR, which counts numbers such as "15.8" or "8000."

5 Discussion

As Table 3 shows, the coverage of the UI corpus by the GSL 1 (level 1) is as low as 61.1%, while other corpora are all above 70%. This means that non-natives with introductory vocabulary knowledge would have difficulty in reading UIs written in English compared with other types of English. The coverage by the GSL 2 (level 2) is almost at the same level, but the coverage by the AWL (level 3) is higher with a wide difference. In short, the UI corpus tends to use higher-level words. The overall coverage (GSL+AWL) either including or excluding variables is close. In Table 4, like the result of GSL+AWL, the coverage of BNC level 1 for the UI corpus is less than 70%, which is lower than other corpora. However, the coverage of level 2 to 4 is higher than others. This tells that the UI corpus includes less introductory level words and more intermediate level words. The overall coverage does not differ much.

As listed in Table 5, the UI corpus shows the least number in both Guiraud and Herdan indices compared with other corpora. This means that the UI corpus has the least diversity. Table 6 illustrates that the lexical diversity of the UI corpus is the least even after adjusting the corpus sizes. Therefore, in all three measures, the UI corpus has the least lexical diversity.

We included the CNN Tech corpus to see if IT-related words affect the results. The coverage and diversity results indicate that the CNN Tech is closer to general English than to UI English, and IT-related words do not affect the characteristic of the text.

When we see Table 7, we successfully pick up characteristic words in the UI corpus. Some words are closely related to software operations (click, configure, etc.), and other words look rather general (add, allow, change, etc.) but are used more frequently in UI English than in general English. We would like to further this study in the future by adding POS information or by extending to top 100, for instance.

6 Conclusion

From the results of the coverage by word lists and the lexical diversity, we conclude that UI English contains more words that are above introductory level but that it is less diverse than general English. Therefore, for non-natives to use software smoothly, they need to learn a relatively limited number of words that frequently appear in UIs, like the ones shown in the result of characteristic words. However, frequently appearing words include advanced level words such as "configure." For non-natives to use software comfortably, developers may avoid using these words in UIs, but at the same time, such words should be taught in English education because a natural language or existing UIs do not change easily in a short period of time.

Acknowledgements. This research report has made use of the English Vocabulary Profile. This resource is based on extensive research using the Cambridge Learner Corpus and is part of the English Profile program which aims to provide evidence about language use that helps to produce better language teaching materials. See http://www.englishprofile.org for more information.

References

1. The Business Software Alliance: Free & Fair Trade, http://www.bsa.org/country/PublicPolicy/global-trade/fair-trade.aspx (accessed on February 11, 2013)
2. del Galdo, E.: Culture and Design. In: del Galdo, E., Nielsen, J. (eds.) International User Interfaces, pp. 74–87. Wiley Computer Publishing, John Wiley & Sons (1996)
3. Esselink, B.: A Practical Guide to Localization. Benjamins, John Publishing Company (2000)
4. Russo, P., Boor, S.: How Fluent is Your Interface?: Designing for International Users. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI 1993, pp. 342–347. ACM Press, New York (1993)
5. Barber, W., Badre, A.: Culturability: The merging of culture and usability. In: Proceedings of the 4th Conference on Human Factors and the Web, pp. 1–14 (1998)
6. Marcus, A., Gould, E.W.: Crosscurrents: cultural dimensions and global Web user-interface design. Interactions 7, 32–46 (2000)
7. Nielsen, J., del Galdo, E., Sprung, R., Sukaviriya, P.: Designing for international use (panel). In: Nielsen, J. (ed.) Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 291–294. ACM, New York (1990)
8. Apple Inc.: Apple Publications Style Guide, https://developer.apple.com/library/mac/documentation/UserExperience/Conceptual/APStyleGuide/APSG_2009.pdf (accessed on February 24, 2013)
9. Microsoft Corporation: Microsoft Manual of Style 4th edn. Microsoft Press, Redmond (2012)
10. DeRespinis, F., Hayward, P., Jenkins, J., Laird, A., McDonald, L., Radziski, E.: The IBM Style Guide: Conventions for Writers and Editors. IBM Press, Upper Saddle River (2011)
11. West, M.: A general service list of English words, with semantic frequencies and a supplementary word-list for the writing of popular science and technology. Longman, London (1953)
12. Coxhead, A.: A New Academic Word List. TESOL Quarterly 34, 213–238 (2000)
13. Nation, P.: A study of the most frequent word families in the British National Corpus. In: Bogaards, P., Laufer, B. (eds.) Vocabulary in a Second Language: Selection, Acquisition, and Testing, pp. 3–13. John Benjamins Publishing Company, Amsterdam (2004)
14. Konstantakis, N.: Creating a business word list for teaching Business English. Elia 7, 79–102 (2007)
15. Millar, N., Budgell, B.S.: The language of public health—a corpus-based analysis. Journal of Public Health 16, 369–374 (2008)
16. Nation, P.: How Large a Vocabulary is Needed For Reading and Listening? Canadian Modern Language Review 63, 59–82 (2006)
17. English Profile, http://www.englishprofile.org/
18. Šišková, Z.: Lexical Richness in EFL Students' Narratives. Language Studies Working Papers 4, 26–36 (2012)
19. Tweedie, F., Baayen, R.H.: How Variable a Constant be? Measures of Lexical Richness in Perspective. Computers and the Humanities 32, 323–352 (1998)
20. Guiraud, P.: Les caractères statistiques du vocabulaire: essai de méthodologie. Presses Universitaires de France, Paris (1954)
21. Herdan, G.: Type-token mathematics: a textbook of mathematical linguistics. Mouton & Co., The Hague (1960)

22. Baayen, R.H.: Analyzing Linguistic Data: A Practical Introduction to Statistics using R. Cambridge University Press (2008)
23. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19, 61–74 (1993)
24. Kilgarriff, A.: Comparing corpora. International Journal of Corpus Linguistics 6, 97–133 (2001)
25. Chujo, K., Utiyama, M.: Selecting level-specific BNC applied science vocabulary using statistical measures. In: Selected Papers from the Fourteenth International Symposium on English Teaching, pp. 195–202 (2005)
26. Android Git repositories, https://android.googlesource.com/ (accessed on July 31, 2012)
27. Microsoft Language Portal, http://www.microsoft.com/Language/en-US/Translations.aspx (accessed on August 01, 2012).
28. Francis, W.N., Kučera, H.: Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (1979), http://icame.uib.no/brown/bcm.html (accessed on January 14, 2013)
29. Ide, N., Fellbaum, C., Baker, C., Passonneau, R.: The manually annotated sub-corpus: a community resource for and by the people. In: Proceedings of the ACL 2010 Conference Short Papers, pp. 68–73. Association for Computational Linguistics, Stroudsburg (2010)
30. Tech category of CNN.com, http://edition.cnn.com/TECH/ (accessed on February 24, 2013)
31. Anthony, L.: AntWordProfiler (Version 1.3.1) (Computer Software) (2012), http://www.antlab.sci.waseda.ac.jp/
32. The R Project for Statistical Computing, http://www.r-project.org/
33. Baayen, R.H.: languageR (Computer Program) (2011), http://cran.r-project.org/web/packages/languageR/
34. Anthony, L.: AntConc (Version 3.3.5) (Computer Software) (2012), http://www.antlab.sci.waseda.ac.jp/