Developing a Corpus Concordancer as a Learning Tool for EST Students



Shoichi Tsuchida

Shi Laboratory Department of Communication Engineering and Informatics The National University of Electro-Communications (UEC Tokyo)

Jie Shi

Graduate School of Informatics and Engineering The National University of Electro-Communications (UEC Tokyo)

Abstract

Corpus tools are essentially tools for linguistic research or treating texts for linguistic analysis, e.g. lexical features, instead of being educational or learning tools for ESP practitioners and students. EFL learners or users who must handle high level academic written or spoken communication in English independently or with little professional assistance tend not to choose a corpus tool as their first choice to solve language or linguistic problems. This research demonstrates the process of developing a new concordancer that is not only an analytical tool but also a learning tool especially for EFL learners to understand professional and genre-based English. Specific stages of the development of the new concordancer will be reported, i.e. the choice of a computer language, selection of functions, interface design, programming, database preparation, user-friendliness features, testing, and piloting. In addition, the result of a survey to find out the preferences of concordancer functions of graduate students at a Japanese university of science and engineering will be presented.

Introduction

Definition of Corpus by Cheng (2012): A corpus is a collection of texts that has been compiled for a particular reasonl. In other words, a corpus is not a collection of texts regardless of the types of texts collected or, if a variety of text types (i.e., genres) are in the corpus, the relative weightings assigned to each text type. A corpus, then, is a collection of texts based on a set of design criteria, one of which is that the corpus aims to be representative. (p. 3)



Classification of Corpora

- Specialised corpus
- General corpus
- Comparable corpora
- Parallel corpora
- Learner corpus
- Pedagogic corpus
- **Historical or diachronic** corpus
- Monitor corpus (Hunston 2002, pp. 14-16)

Selection of 100 words from BNC list Verb: be, have, do, TOP get, go, say, know, think, 100 see, come, mean, want, take, look, make, put, give Auxiliary : will, can would, could Conjunction : and, that, but, if or, as, when because, cos 57457 types Pronoun: 12, Preposition: 12 Adverb: 18, Noun: 6 10 million words

Figure 1.1 Selection of 100 words from the BNCspoken list (Tono, 2011c, p. 7)

Research Purpose

To develop a corpus linguistic tool which is user-friendly for researchers, teachers and leaners especially of EST.

Methodology & Design

A. Programming Language: Python

All Abstract Definitions	📀 Corpus tools are used to analyze and process text data to make corpora.							
O Discussion of results O Literature review	First, you must choose a file that you want to analyze or process from left side menu.							
 Methodology section Problem-solution pattern Deference to courses 	🔇 The file should be chosen from your own file or Shi lab. Data Base.							
• Tables, graphs or figures	✓ However, the function "Charts" doesn't require a particular file.							
<u>Academic Position/Role:</u> All Junior Faculty 	Next, select a function that you want to use.							
 Junior Graduate Student Junior Undergraduate Post-doctoral Fellow 	If you want to use download version, click following button "Download".							
© Researcher © Senior Faculty © Senior Graduate Student © Other	使い方							
Native speaker status:	🤣 コーパスツールとはテキストデータを分析、もしくはコーパスを作るためのものです。							
O Non-native Speaker O Near-native Speaker	🤣 まずは、左側のメニューから処理を行いたいファイルを選択します。							
Native Speaker	✓ 手元のファイルか史研究室データベースから選択してください。							
Other Data-Bases	✓ 但し、チャート機能(Charts)を使う場合はファイルの選択は必要ありません。							

KWIC



Word List



	Content:	🧭 Corpus tools are used to analyze and process text data to make corpora.
	In the treasure hunting process, the hunters could only get few hints with simple words, therefore they must think over with the experience they learned before. The goal of this paper is to propose a treasure hurting learning	🧭 First, you must choose a file that you want to analyze or process from left side menu.
		🧭 The file should be chosen from your own file or Shi lab. Data Base.
	model and implement a system based on the model. Within the treasure hunting learning model students will become a treasure hunter when they are learning	✓ However, the function "Charts" doesn't require a particular file.
	in the field. The learning model makes learning more fun by asking students some questions according to their	Next, select a function that you want to use.
le o tr n t	learning records and physical position. In order to make the system portable and treasure hunting liked, we implement the treasure hunting learning model with the and simple and common cellphone, that is the text-mode cellphone. To verify the learning model is working, a traditional history and culture course in college was token as the experiment.	✓ If you want to use download version, click following button "Download".
		使い方
		🤣 コーパスツールとはテキストデータを分析、もしくはコーパスを作るためのものです。
		🤣 まずは、左側のメニューから処理を行いたいファイルを選択します。
		✓ 手元のファイルか史研究室データベースから選択してください。
		✓ 但し、チャート機能(Charts)を使う場合はファイルの選択は必要ありません。

N-grams (N=2)







Figure 4.1 Number of top 39 U.S. computer science departments that use each language to teach introductory courses

B. Aiming at User-Friendliness

- **GUI = Interface**
- Levels or Fields
- Display of Functions
- **Bilingual instructions**
- Participatory features

C. Selection of Functions

- KWIC (Key Word In Context)
- Word List
- N-gram
- Cluster
- Collocate
- File View
- Charts
- Exercises
- Participatory

File View



ollocates												
Change this File	LIEC	KWIC Word Lis Participatory	LEVELS st N-grams	FI Clusters C	IELDS Collocates	FileView C	harts Excercise	e				
Alex Chang269676-2006- A_Treasure_Hunting_Learning_A	Colloc O Display	ates what word is used	<u>"learn</u> with particulo	ing" ar word								
Content:	✓ 特定の単語と共起されている単語を頻度順に表示											
In the treasure hunting process, the			Word	Frequenc	v Frequency(L	.) Frequency(R	0					
nunters could only get tew hints with	Word		model	5	0	5	<u></u>					
over with the experience they learned	learning		hunting	3	3	0	-					
before. The goal of this paper is to			The	1	1	0	-					
propose a treasure hunting learning			are	1	1	0	-					
the model Within the treasure hunting	OK		in	1	0	1						
learning model students will become a			makes	1	1	0						
treasure hunter when they are learning			more	1	0	1	-					
in the field. The learning model makes			records	1	0	1						
some questions according to their			the	4	1	0						
learning records and physical position. In			thoir	4	1	0						
order to make the system portable and			uiell			U						
treasure hunting liked, we implement the												
treasure hunting learning model with the												
is the text-mode cellphone. To verify												
the learning model is working, a												
traditional history and culture course in												
college was taken as the experiment.												

Conclusion

- Have developed a concordancer for students or researchers of EST
- Python: A suitable language for interface and lowmaintenance
- **Positive feedback from graduate students**
- 6 functions completed
- Web-browser style of interface

References

Cheng, W. (2012). Exploring Corpus Linguistics. New York, United State of America: Routledge.

Guo, P. (2014, 7 7). Python is Now the Most Popular Introductory Teaching Language at Top U.S. Universities. Retrieved from Communications of the ACM: http://cacm.acm.org/blogs/blog-cacm/176450-python-is-nowthe-most-popular-introductory-teaching-language-at-top-us-universities/fulltext

Hunston, S. (2002). Corpora in Applied Linguistics. Cambridge, United Kingdom: Cambridge University Press. McCarthy, M., & O'Keeffe, A. (2010). The Routledge Handbook of Corpus Linguistics. Routledge .

Nesi, H. (2013). ESP and Corpus Studies. In B. Paltridge, & S. Starfield, The Handbook of English for Specific *Purposes* (p. 407). Wiley-Blackwell.

Tono, Y. (2011). TaLC in action: recent innovations in corpus-based English language teaching in Japan. In A. G. Frankenberg, G. Aston, & L. Flowerdew, New Trends in Corpora and Language Learning (pp. 7-8).

Bloomsbury.