4-18-2016

Frequently Used Phrases in China's National Newspaper.

Randy Josleyn

Frequently Used Phrases in China's National Newspaper.

Abstract

Corpora are excellent resources for learning, particularly considering research showing the importance of frequently used word clusters (called lexical bundles or collocations) in promoting learner fluency. However, in the context of Chinese language, most of the available corpus resources seem to be unrepresentative of how native Chinese use language in everyday life, possibly due to the influence of writers' awareness of censorship on the Chinese internet.

The goal of this endeavor was to create a corpus of reliably natural text from China's national newspaper, The People's Daily（人民日报，人民网），for the purpose of identifying lexical bundles that serve to create structure in Chinese sentences in the news register. The corpus was extracted from The People's Daily website using a web crawler, to a total of more than five hundred articles and about one million Chinese characters.

As such, the presentation will reveal several trends in collocation, which can serve as a resource to develop learning materials to improve Chinese language learning, especially for improving Chinese reading skill in the domain of news articles.

乔智乐

Lexical Characteristics of Chinese News Media

Randy Josleyn

2016年4月18日

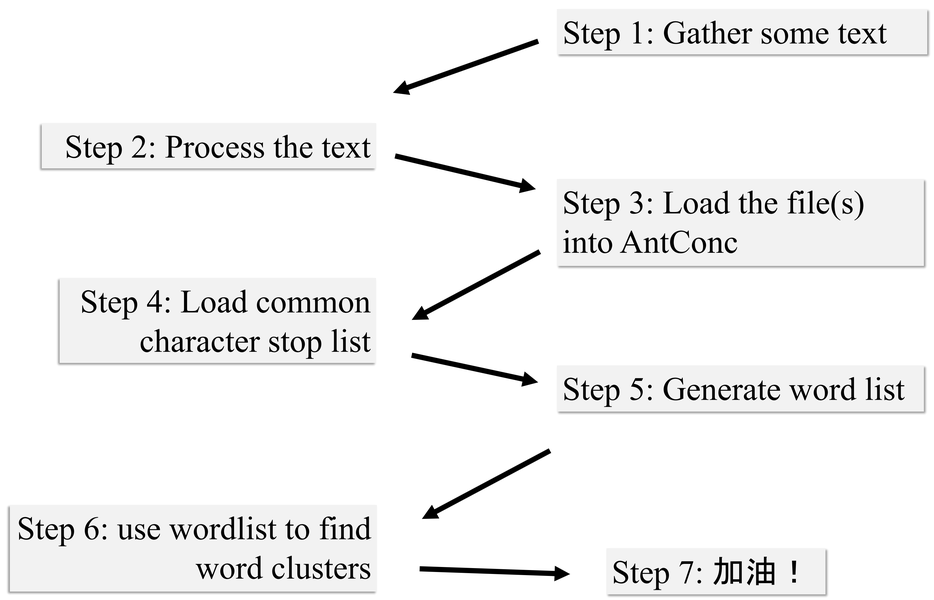
47

What are the linguistic patterns of Mandarin Chinese news? This research aimed to uncover some of the common features of Chinese writing in order to help Chinese learners learn how to read.

Word Frequency

Teaching often focuses too heavily on character knowledge, altogether overlooking words. Try searching the internet for “Chinese most common words,” and all you will get are character lists. Here are some of the most common words to appear in news articles in this Qdaily corpus. The basic technique for this method is described on the right-hand side.

Rank	Frequency	Word	Definition
1	684	公司	company
2	403	苹果	Apple (the company)
5	290	电影	movie
10	241	产品	product
11	239	手机	cell phone
12	234	游戏	game
14	222	亿	100 million
19	200	市场	market
22	193	用户	user
24	185	万	10 thousand
25	181	迪士尼	Disney
28	158	投资	invest
29	156	合作	cooperate
31	150	腾讯	Tencent (makers of QQ)
34	140	发布	announcement
36	136	目前	at the present
39	132	小米	Xiaomi (a Chinese company)
40	130	应用	use
43	128	服务	service
44	126	来自	come from



What Is a Corpus?

A corpus is a collection of texts and/or spoken language. It is used by linguists to look for statistical language patterns that would be hard to find by hand. (The word cloud above is one example of the use of a corpus.)

The Project Corpus

The corpus for this research was compiled by pulling text from the Chinese news and entertainment magazine Qdaily. Each article was copied into text files over several weeks. Since Chinese words do not have spaces between them, a segmenting program was used to automatically add them. Then, the freeware corpus analysis program AntConc was used to search for patterns in the text.

Background

- Corpora are excellent learning resources, as used in *A Frequency Dictionary of Mandarin Chinese* by Xiao, Rayson, and McEnery
- Existing Chinese corpora are hard to access, creating a need to create one from scratch
- Taguchi (2013) demonstrates how important formulaic language is to promoting learner fluency
- Commonly available learning resources do not account for register (context) differences