Independent Study: Character Frequency Analysis for Simplified Chinese Learners

**INTRODUCTION**

Mandarin is the most commonly spoken language in the world, representing roughly a billion speakers and 14% of the world population (1). Additionally, since 1979, China's economy is the fastest growing in the world and has the second-largest GDP of all countries (behind only the US) (2). Considering China's surging world presence, the opportunity to learn Mandarin and the Simplified Chinese written language has never been more valuable.

For native English speakers, the prospect of learning the Mandarin language presents significant challenge. English and Mandarin originate from entirely independent language families, with different syntax, orthography, and phonetic sound inventories(3). The most significant pain point for would-be Mandarin learners is the lack of educational resources. The most popular language learning software, Duolingo, has won numerous awards over its 4 year life. It offers thorough language learning curriculums for 13 languages, but not Mandarin(4). Prospective Mandarin learners have few places to turn to apart from university courses, which present their own downsides. Indeed, most Mandarin learners are university students with various majors and interests. In these cases, communicating their specific knowledge using Mandarin is particularly difficult.

Modern language learning techniques involve some amount of word frequency analysis to determine which words are most prudent to learn. Pronouns (he, she, it), articles (the, a, your), conjunctions (for, and, but), and prepositions (of, on, at) inevitably top these lists in any language. Such words are important as they provide the building blocks of language, but are not enough to communicate beyond very basic sentences. College-level courses require a higher level vocabulary specific to each field of study.

**ABSTRACT**

This study aims to analyze academic sources of several popular college majors to provide Simplified Chinese word frequency statistics unique to each major. These results will provide students who wish to study abroad the vocabulary they need to communicate within their major.

**METHODOLOGY**

An important distinction of the written Chinese language is that two character sets exist, the Simplified and the Traditional. This study will focus on the more concise Simplified character set, as it is more approachable for a beginner.

To break our problem of analyzing Chinese language word frequency, we can first take a high-level approach to doing the same for the English language. English words consist of a series of letters delimited by spaces and punctuation. These spaces can easily be used as boundaries to tell our software where a word begins and ends. The process proceeds in a straightforward fashion as we iterate over words and tally them to create our word frequency statistics.

Translating this process to Chinese proves difficult. Except in Romanized Pinyin, the Chinese language has no word markers equivalent to the English 'space'. Word ambiguity becomes an obstacle, with the only solution being to create a Chinese dictionary and to make “best guess” distinctions in ambiguous scenarios. Such is an obstacle for modern Chinese character typing and processing programs today and is beyond the scope of this study(5).

For Simplified Chinese learners, recognizing individual characters and their meanings will provide a firm and sufficiently challenging foundation for learning the broad vocabulary required in any field. To use mathematics as an example:

# The character数, meaning “number” is the most frequent and inarguably most important character. It also forms the base of the words 数学 (integer), 整数 (natural number), and 分数 (fraction), among numerous others.

Counting each character individually will allow learners to know which characters are most immediately necessary in their field of study. Since the Simplified Chinese character set contains 6,500 characters, this knowledge is tremendously useful, and also applies to the Traditional Chinese character set (6).

Upon deciding to tally individual characters, the next decision to be made is what form of scholarly articles to analyze. The obvious choice research papers, which overwhelmingly take the form of PDF documents.

The unfortunate truth is that automating the analysis of PDF documents is a monumental task. Adobe goes to great lengths to make such analysis as difficult as possible, providing password, file protected, permission, and encryption services(7). By some estimates, Adobe Acrobat, the program for creating and viewing PDF files, is over 13 million lines of code(8). Creating even a rudimentary version of this program for extracting simple text is an enormous project and one which I spent weeks on with little success.

In leiu of PDF files, the Internet provides an easy and straightforward source of academic information. In Simplified Chinese, the most prominent source of educational content is Baidu's [Baike](http://baike.baidu.com/) (encyclopedia). It can be thought of as a Chinese alternative to the Wikipedia.

Using the Internet as a foundation provides several benefits, namely: the results are easy to replicate and the results of this study an easily be expanded.

Upon initial analyses of character frequency Baidu's online encyclopedia articles, it was clear that the data required scrubbing. Tremendous overlap existed between articles of remarkably different disciplines. To get the data at the heart of this study, characters representing punctuation, pronouns,articles, prepositions, and conjunctions were removed. Additionally, all of the top 100 Chinese characters were removed to eliminate redundancy, as these are characters learners of any discipline will undoubtedly need to know.

For the generation of statistics, I chose one major from each of the two main categories of disciplines. Anthropology from the social sciences and Mathematics from the STEM sciences. Additionally, I chose Computer Science, my own field of study. Math and Anthropology provide a contrast with which to grade the results of our data, while Computer Science provides a foundation for my own potential journey into learning the Simplified Chinese character set.

It is worth noting that this study inspired a similar project in which I collaborated with another student to create a website capable of analyzing word frequency of the English language in DOCX documents. [Essaystats.com](http://essaystats.com/) is a free word analysis utility with graphing functions and login support to allow analyses to be saved. It's an interesting and fun offshoot of this study.

To produce that data, ten topics of each major/field were chosen for analysis.

ANTHROPOLOGY

http://baike.baidu.com/item/%E4%BA%BA%E7%B1%BB%E5%AD%A6/398754 - 人类学 - Anthropology

http://baike.baidu.com/view/200664.htm - 体质人类学 - Physical Anthropology

http://baike.baidu.com/view/744379.htm - 民族志 - Ethnography

http://baike.baidu.com/view/2294365.htm - 语言人类学 - Linguistic Anthropology

http://baike.baidu.com/subview/51552/14970898.htm - 认识论 - Epistemology

http://baike.baidu.com/subview/85866/14945582.htm - 文化人类学 - Cultural Anthropology

http://baike.baidu.com/view/3586780.htm - 自然人类学 - Social Anthropology

http://baike.baidu.com/view/2673780.htm - 象征人类学 - Symbolic Anthropology

http://baike.baidu.com/view/7793278.htm - 诠释人类学 - Interpretation of Anthropology

http://baike.baidu.com/view/587742.htm - 语源学 - Etymology

MATHEMATICS

http://baike.baidu.com/subview/1284/13645654.htm - 数学 - Mathematics

http://baike.baidu.com/view/678110.htm - 数量 - Quantity

http://baike.baidu.com/subview/160039/8092759.htm - 结构 - Structure

http://baike.baidu.com/subview/507074/8050350.htm - 变化 - Variety

http://baike.baidu.com/subview/31260/9223598.htm - 空间 - Space

http://baike.baidu.com/view/1527.htm - 信息 - Information

http://baike.baidu.com/view/17541.htm - 算术 - Arithmetic

http://baike.baidu.com/view/2153944.htm - 基础数学 - Basic Mathematics

http://baike.baidu.com/view/556393.htm - 代数学 - Algebra

http://baike.baidu.com/view/15136.htm - 几何 – Geometry

COMPUTER SCIENCE

http://baike.baidu.com/subview/92404/5815703.htm - 计算机科学 - Computer Science

http://baike.baidu.com/view/3314.htm - 计算机 - Computer

http://baike.baidu.com/view/629336.htm - 计算机系统结构 - Computer Architecture

http://baike.baidu.com/subview/37/6030295.htm - 软件 - Software

http://baike.baidu.com/view/8332.htm - 程序设计 - Programming

http://baike.baidu.com/view/689084.htm - 理论计算机科学 - Theoretical Computer Science

http://baike.baidu.com/view/920697.htm - 计算理论 - Theory of computation

http://baike.baidu.com/view/15076.htm - 信息论 - Information Theory

http://baike.baidu.com/view/1438395.htm - 编码理论 - Coding Theory

http://baike.baidu.com/subview/1659/7112047.htm - 软件工程 - Software Engineering

**RESULTS – MATHEMATICS**

数 433 %5.4105 Number

科 116 %1.4495 Family/Branch

论 92 %1.1496 Theory

理 87 %1.0871 Reason

分 84 %1.0496 Minute/Fraction

辑 80 %0.9996 Edit

量 69 %0.8622 Amount

代 68 %0.8497 Generation/Age

编 67 %0.8372 Compile/Arrange

现 61 %0.7622 Current/Present

百 57 %0.7122 Hundred

研 55 %0.6872 Research

究 55 %0.6872 Investigate

展 54 %0.6747 Show/Exhibit

几 51 %0.6373 A few, how many

解 47 %0.5873 Solution

等 46 %0.5748 Wait

关 46 %0.5748 Shut/close

词 46 %0.5748 Word/term

定 45 %0.5623 Set/order

**RESULTS – ANTHROPOLOGY**

人 658 %6.0278 People

类 455 %4.4356 Class

体 217 %2.1154 Body/form

研 197 %1.9205 Research

究 195 %1.9010 Investigate

质 155 %1.5110 Quality

科 131 %1.2771 Family

现 124 %1.2088 Current

种 123 %1.1991 Kind/type

展 97 %0.9456 Show

进 89 %0.8676 Enter/advance

与 86 %0.8384 Versus

物 74 %0.7214 Object/thing

等 74 %0.7214 Wait

代 71 %0.6921 Generation

论 68 %0.6629 Theory

古 66 %0.6434 Ancient

社 65 %0.6337 Society

编 62 %0.6044 Compile

族 61 %0.5947 Clan

**RESULTS – COMPUTER SCIENCE**

机 735 %3.5063 Machine

计 714 %3.4062 Meter/Plan/Idea

算 705 %3.3632 Count

科 257 %1.2260 Family

器 253 %1.2069 Device

数 246 %1.1736 Number

理 243 %1.1592 Reason

程 236 %1.1258 Journey

电 234 %1.1163 Electric

系 233 %1.1115 System

统 225 %1.0734 Series

件 187 %0.8921 Item

工 174 %0.8301 Work

分 153 %0.7299 Minute

现 150 %0.7156 Current/Present

展 147 %0.7013 Show/exhibit

存 140 %0.6679 Deposit

制 139 %0.6631 Make/manufacture

设 135 %0.6440 Assume

信 131 %0.6249 Letter

性 129 %0.6154 Sex/gender

**CONCLUSION**

Among the data, little overlap exists between fields. Even between mathematics and computer science, highly related field, few words are seen in common among highest frequencies. Especially among the top 5 words of each field, the data seems to represent the language at the heart of each respective field. Most notably, the singly highest frequency character for each data set is particularly representative:

Anthropology: 人 - People

Mathematics: 数 - Number

Computer Science: 机 – Machine

It seems as if we've found the core language that sets each field apart, in addition to developing an algorithm to find such language. These characters are not found among the overall top 100 characters of the Simplified Chinese character set. For a written language as complex as Chinese, where some characters contain as many as twenty strokes, focusing your attention on the characters you will see most often is paramount. Studying one of these resulting sets of characters in addition to the overall top 100 characters would be an ideal jumping off point for a student studying abroad in China.

**WORKS CITED**

1.Mandarin Chinese https://en.wikipedia.org/wiki/mandarin\_chinese (accessed Aug 2015).

2. China https://en.wikipedia.org/wiki/china (accessed Dec 2015).

3. The differences between English and Chinese http://esl.fis.edu/grammar/langdiff/chinese.htm (accessed 2015).

4. Duolingo | Course Offerings https://www.duolingo.com/courses/en (accessed 2015).

5. Wong, K.-F.; Li, W.; Xu, R.; Zhang, Z.-S. Introduction to Chinese Natural Language Processing. *Synthesis Lectures on Human Language Technologies* 1–148.

6. Simplified Chinese Characters https://en.wikipedia.org/wiki/simplified\_chinese\_characters (accessed 2015).

7. Adobe Acrobat DC https://acrobat.adobe.com/us/en/how-to/pdf-file-password-permissions.html (accessed 2015).

8. Hey Adobe, hackers have access to your Source Code! http://labs.bromium.com/2013/10/14/hey-adobe-hackers-have-access-to-your-source-code/ (accessed 2015).