The Evolution of Written Vernacular Chinese

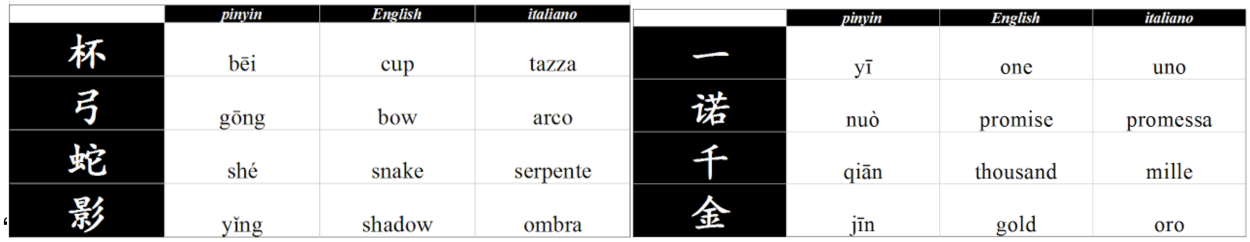
Yachen Sun

Background

Despite the fact that Chinese is one of the oldest written languages in the world with at least six thousand years of history, modern written Chinese is still a very young language that has undergone some significant changes in the last century. In this essay, we will explore some of the changes that has happened, is happening, or will happen to the modern written Chinese language.

During the most part of the Chinese history, Classical Literary Chinese, or *wenyan*, has been the language used to write Chinese texts. However, starting from the Qin Dynasty (221 BC), the spoken Chinese language used by most people in the society has gradually diverged from the written language. There are several potential reasons behind this divergence. First, Classical Literary Chinese itself is a relatively vague language with a focus on implied meaning and abbreviated forms, and the readers very often need to fill in grammatical and semantic relationships between words and phrases using their own intuition. In other words, the interpretation of Classical Literary Chinese often involves a fair amount of uncertainty, and also requires high degrees of education and effort from the readers. Figure 1 gives two example idioms from Classical Literary Chinese. 1(a) summarizes a story with four noun characters, and 1(b) describes a virtue using two noun phrases. We can see from these examples that grammatical structures are often abbreviated or ignored in Classical Literary Chinese, and the users often need to make allusions to other literary sources in order to understand certain expressions. This property of the Classical Literary Chinese language is largely caused by the fact that paper was not invented until around 105 AD[1]. Before paper was invented, written Chinese were mostly carved onto media such as turtle shells, stone plates, or bamboo slips. The carving process is time consuming, which means that people would try to compress as much information into a single character as possible. This largely led to the vague and abbreviated forms of Classical Literary Chinese.

Another factor that plays an important role in the separation of spoken Chinese from written Chinese is politics and education[2]. Classical Literary Chinese is taught through repetitive copying and recitation of classical texts with little guided explanation about what these texts actually meant. The students were expected to gradually develop understandings of the texts on their own, which often required them to possess certain cultural background or experiences. Not only was Classical Literary Chinese a language not accurate and descriptive enough to be used in every day life, but the barrier to mastering the language was also pretty high. Thus, the spoken Chinese language has gradually diverged from the written Chinese language.



	pinyin	*English*	*italiano*
杯	bēi	cup	tazza
弓	gōng	bow	arco
蛇	shé	snake	serpente
影	yǐng	shadow	ombra

(a)

	pinyin	*English*	*italiano*
一	yī	one	uno
诺	nuò	promise	promessa
千	qiān	thousand	mille
金	jīn	gold	oro

(b)

Figure 1. Idioms from Classical Literary Chinese. *Each diagram contains the single character translations of each character in the idiom into three other languages. (a) This idiom is a metaphor for people who are overly suspicious or paranoid about nonexistent things/*

situations. The four characters are actually a summarization of a story in Feng Su Tong Yi: Guai Shen[3]*: Ying Bin invites Du Xuan over for a drink. Du Xuan saw the shadow of a bow hanging on the wall in his cup, and he suspected that he might have swallowed a snake during drinking. (b) This idiom means that someone takes the promises that he/she makes very seriously.*

During the Ming and Qing dynasties (1368-1912), authors started to compose novels in vernacular Chinese. In fact, the Four Great Classical Novels of Chinese Literature were all written during this era, and all of them were composed in vernacular Chinese. Nevertheless, Classical Literary Chinese continue to dominate formal writing. After China became a republic (1912), many progressives started to see Classical Literary Chinese as hindering the education and literacy in China[3]. Two language movements during this era greatly shaped the written language used in mainland China today: the *Baihuawen* (Written Vernacular Chinese) Movement and the Simplified Chinese Movement. The Baihuawen Movement mainly pushed for vernacular Chinese to become the major form of literary expression for Chinese people. The Simplified Chinese Movement encouraged people to write with simplified versions of traditional Chinese characters. These movements have greatly changed how people write and speak today in mainland China, and led to the production of a vast number of literary works in vernacular Chinese.

The Past, Now, and Future of Written Vernacular Chinese

Written vernacular Chinese is a relatively young language that is heavily influenced by spoken Chinese, but at the same time retaining many morphemes and expressions from Classical Literary Chinese[4]. Besides spoken Chinese and Classical Literary Chinese, I here propose four major factors that have shaped and will continue to shape the evolution of Written Vernacular Chinese: migration and mixing of populations, the Internet, westernization and the "translation tone," and typing habits, with a special emphasis on the third factor - westernization.

Migration and Mixing of Populations

China has seen extensive internal migration since the implementation of the Reform and Opening-Up Policy ("改革开放"). According to Chan and Bellwood (2011), " China's urban population has grown by about 440 million to 622 million in 2009. Of the 440 million increase, about 340 million was attributable to net migration and urban reclassification. Even if only half of that increase was migration, the volume of rural-urban migration in such a short period is likely the largest in human history."[5]

Such massive rural-urban migration will likely lead to the mixing of people speaking different dialects, and may result in certain morphemes and expressions unique to particular dialects becoming more and more accepted by the common spoken Mandarin, and eventually making its way into Written Vernacular Chinese. For example, the word "给力" (pronounced as "gei li;" the characters mean "give force") is an expression used in the Min Dialect (闽南语) spoken at Zhangzhou, Fujian Province to describe something as exciting or awesome[6]. Since Year 2010, this word has been very frequently used in social networks and online forums, especially in discussions about the World Cup that year. As the word became more and more popular and accepted by people throughout mainland China, on November 10th, 2010, *People's Daily*, the Communist Party's mouthpiece newspaper, used "gei li" in its headline title.

Using neighbor-net analysis, Hamed (2005) identified "strong homogenization forces related to diglossia and heavy borrowing" among various dialects in China. In particular, Mandarin, from which the standard spoken language in China is derived, has weak dialect boundaries and "is virtually related to all dialects."[7] Such homogenization forces, together with the fast spread of new words and language phenomena over the Internet, will very likely continue to allow vernacular Chinese to "borrow" words and expressions from regional dialects. Unfortunately, there are currently very few quantitative studies analyzing how words from dialects makes their way into vernacular Chinese.

The Age of the Internet

Similar to English, Chinese has been changed by the use of the Internet as well. Various Internet slang words and emoticons have made their way into spoken and written vernacular Chinese, such as "gei li", which was mentioned in the last section. The wide usage of microblogging sites like Weibo and Twitter has also led people to adopt hashtags and pithier expressions. One particularly interesting phenomenon unique to Chinese Internet slangs, though, is the popular usage of puns.

Xiao and Link (2013) wrote in the Wall Street Journal that due to the heavily monitored environment of China's cyberspace, the Chinese "netizens" had adopted "coded language and metaphors to avoid outright censorship."[7] For example, since "government" is often a censored word that cannot appear in online posts and publications, people often refer to the government as "天朝," which means "heavenly dynasty." Another example is that people also use "真理部,"

or “the Ministry of Truth” in English, to refer to the propaganda department of the Communist Party, which alludes to George Orwell’s famous anti-utopian novel, *1984*. These puns and metaphors allow people to talk about censored entities online with a mischievous hint of sarcasm.

Nevertheless, the usage of Internet puns and metaphors actually extend way beyond censored words and entities. For example, many people use “粉丝” (pronounced as “fen si”) to address someone’s fans, because its pronunciation is very similar to “fans.” Many also use “恐龙” (pronounced as “kong long;” means dinosaurs) to refer to people with horrifying appearances. Thus, besides censorship, many Internet puns and metaphors may be created just for fun. Indeed, we can trace similar phenomena back to Classical Literary Chinese, where metaphors and allusions are also prevalent. The pun and metaphor game is likely a cultural phenomenon that persists in the Chinese language regardless of the exact form of the current language used.

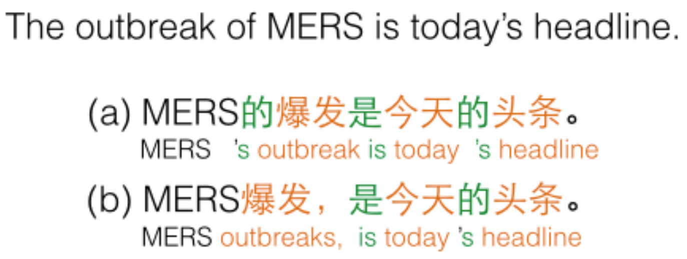
Typing Habits

Most people in mainland China now uses Pinyin input methods to type Chinese characters. In this method, the users type in the pronounciation of the characters to be typed, and then select the desired characters from the list of characters sharing the same pronunciation. Because inputting the pronunciation of one character usually brings up a long list of characters to choose from, most Pinyin input methods now offer a function where the user can input the pronounciation of a phrase or expression, which drastically reduces the possible candidate character combinations. Thus, it is more efficient to type Chinese in phrases rather than in single characters. To this day, no study has investigated the influence of such typing habits on the evolution of the Chinese language, but an interesting hypothesis to pursue is that such typing habits will lead to the preferred usage of multi-character words/expressions over single character words.

Westernization

Yu Guangzhong, a famous author and linguist in China, wrote several papers claiming that one important crisis that the Chinese language is facing is the westernization of Chinese, especially the influence of English on Chinese.

Yu Guangzhong mentioned several different ways in which the modern Chinese language is influenced by western languages, and one particularly interesting point is the use of abstract ideas as subjects and the weakening of verbs[8]. Traditionally, Chinese speakers likes to use objects, expressions, or clauses as the subjects of sentences, whereas English speakers also like to use abstract ideas and noun phrases derived from verb phrases as subjects. Using verb-derived nouns lead to the weakening of verbs. As a result, instead of directly using verbs as the verb in the sentence, English speakers sometimes use a combination of a “weaker” verb and the noun form of the original verb instead. For example, instead of saying “press something,” and English speaker can also say “apply pressure to something.” Recently, such usage has appeared in Chinese as well. An example is demonstrated in Figure 2.



The outbreak of MERS is today’s headline.

(a) MERS的爆发是今天的头条。
MERS ’s outbreak is today ’s headline

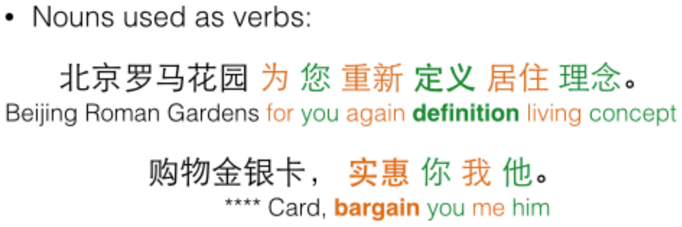
(b) MERS爆发，是今天的头条。
MERS outbreaks, is today ’s headline

Figure 2. Demonstration of the influence of English on Chinese. *Both (a) and (b) are Chinese sentences expressing the same meaning. (a) is the version showing influence by*

English, where the verb “爆发” (outbreak) is used as a noun, and the abstract idea “MERS’s outbreak” serves as the subject of the sentence. (b) is the more natural and original version, where the clause “MERS爆发” (MERS outbreaks) serves as the subject of the sentence.

The phenomenon Yu Guangzhong pointed out is a part of the problem of “verbal nouns.” The analysis of Chen (2005) on the influence of English on Chinese expanded the problem of verbal nouns to the transformations between various part-of-speech (POS) components[9]. English, like all other Latin languages, is an alphabetic language. In such languages, changing the POS of a word usually involves adding or subtracting various prefixes and suffixes from the original word. Chinese, however, is a logographic language, where characters cannot be modified by prefixes and suffixes. Chen (2005) argues that large volumes of Chinese texts translated from English have created a trend in modern Chinese to use variable POS for words. Due to the logographic nature of the Chinese language, instead of modifying the original verb to fit different POS, the same word will be used for various POS, and the specific POS will be decided according to the context. Chen (2005) proposes several particular categories of POS transformations in Chinese as heavily influenced by English translations, including nouns used as verbs, nouns and verbs used as adjectives, and verbs and adjectives used as nouns. Figure 3 shows some examples in each category.

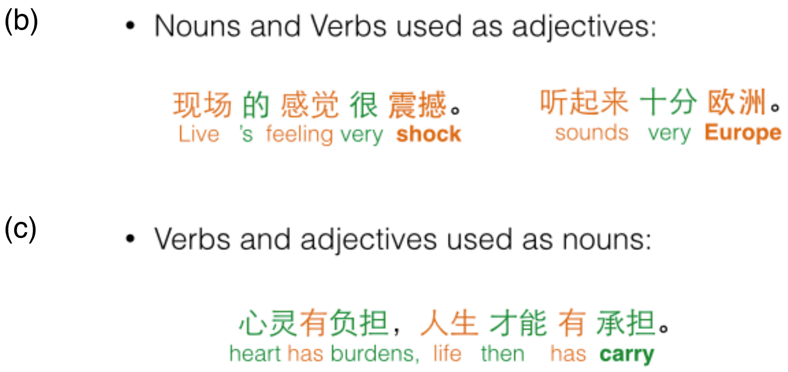
(a)



- Nouns used as verbs:

北京罗马花园 为 您 重新 定义 居住 理念。
Beijing Roman Gardens for you again **definition** living concept

购物金银卡， 实惠 你 我 他。
**** Card, **bargain** you me him



(b)

- Nouns and Verbs used as adjectives:

现场 的 感觉 很 **震撼**。
Live 's feeling very **shock**

听起来 十分 **欧洲**。
sounds very **Europe**

(c)

- Verbs and adjectives used as nouns:

心灵有负担，人生 才能 有 **承担**。
heart has burdens, life then has **carry**

Figure 3. Chinese POS transformations that are influenced by English. *In these examples, the bold fronted characters / phrases are used as a different POS from their traditional usage. In (a), "定义" (definition) and "实惠" (bargain) are traditionally used as nouns, but are used as verbs in the sentences shown. In (b), "震撼" (shock) is typically used as a verb, and '欧洲" (Europe) a noun, but here they are used as adjectives. In (c), "承担" (carry) is a verb, but it is used as a noun in this sentence (it means "responsibility" here).*

Nevertheless, it is unclear whether such POS transformations are really borrowed from English. It is possible that the property where the same word can have different POS depending on context is an intrinsic property of the Chinese language. Since Classical Literary Chinese is a language that focuses on implied meaning rather than grammatical relationships, it is fairly common to use the same character or phrase as different POS in different contexts.

Westernization - the Hypothesis

If variable POS associated with one word in vernacular Chinese is a feature learned from English, we should be able to see increasing instances of such usage over the last century, as China becomes more and more exposed to the western world.

Westernization - Methods and Materials

In order to compare the frequencies of variable POS in vernacular Chinese at different time points in the last century, I selected 9 classic and popular novels that are widely known and praised in contemporary China, with publishing dates ranging from 1926 to 2014. Since novels are heavily influenced by writers' personal styles, I set up a second group of comparisons with random articles drawn from the People's Daily Database[10]. The articles are drawn from years 1947 and 1948, 1981, 2000, and 2014.

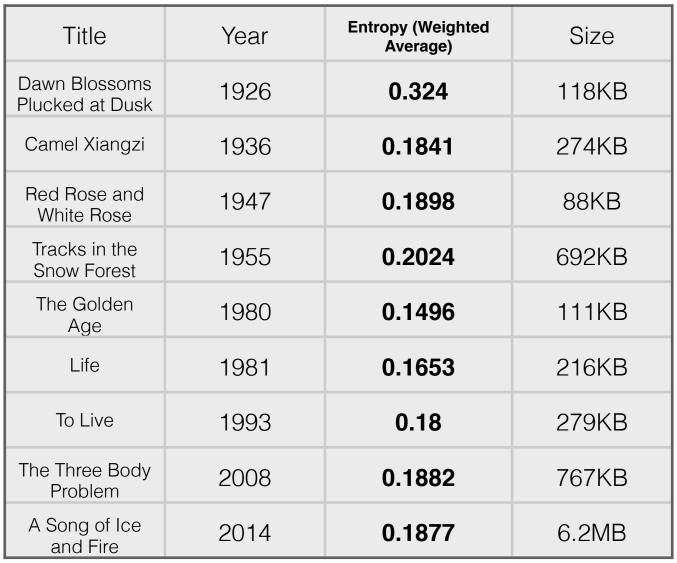
These texts are then processed with the Stanford Chinese Natural Language Parser. Specifically, I used the Stanford NLP segmenter to segment the texts into character phrases[11]. The model used for segmentation is the Chinese Penn Treebank standard. I then used the Stanford NLP POS tagger to tag the POS of each character phrase generated by the segmenter[11]. The model used for tagging is Chinese - DistSim.

The extent to which variable POS is used within each text is evaluated using the Shannon Index, also known as the Shannon Entropy. The formula is shown below:

$H' = -\sum_{i=1}^{R} p_i \ln p_i$ where H' is the Shannon Entropy for a particular word in the text; pi is the proportion of a particular POS compared to all of the occurrences of the word within the text; R is the total number of different POS that the word takes in this text. The Shannon Entropy is a classic metric to measure diversity, as it accounts for both richness and evenness of the categories covered. Thus, the higher the Shannon Index is for a particular word, the more flexible this word is used in terms of POS within this particular text. The overall flexibility in POS usage is then calculated as the weighted average of the Shannon Index of all words with at least two different POS. The weight is the total number of occurrences of the particular word. so that common words are weighted more importantly than uncommon words.

Westernization - Results and Discussion

ANALYSIS OF CLASSIC AND POPULAR LITERATURE



Title	Year	**Entropy (Weighted Average)**	Size
Dawn Blossoms Plucked at Dusk	1926	**0.324**	118KB
Camel Xiangzi	1936	**0.1841**	274KB
Red Rose and White Rose	1947	**0.1898**	88KB
Tracks in the Snow Forest	1955	**0.2024**	692KB
The Golden Age	1980	**0.1496**	111KB
Life	1981	**0.1653**	216KB
To Live	1993	**0.18**	279KB
The Three Body Problem	2008	**0.1882**	767KB
A Song of Ice and Fire	2014	**0.1877**	6.2MB

Table 1. Weighted average entropy of 9 classic and popular novels in vernacular Chinese.

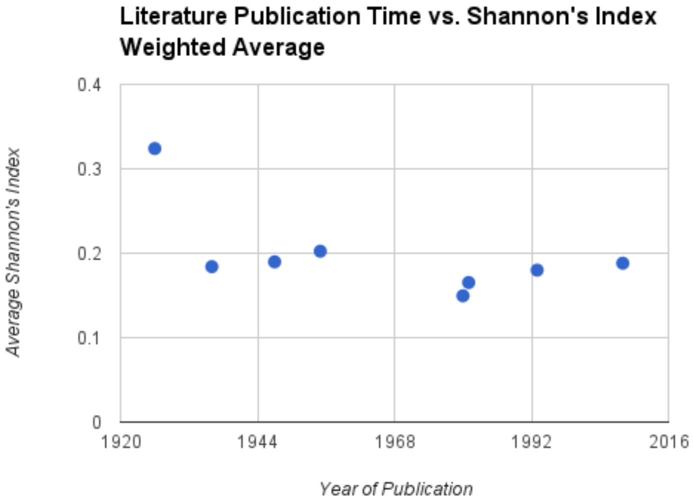


Figure 4. The publication time of novels and the weighted average Shannon Indices.

The weighted average entropies of classic and popular literature are shown in Table 1 and Figure 4. There does not seem to be an obvious relationship between publication times and Shannon Indices, which could mean the lack of correlation between the two, or that the methodology used fails to capture the correlation, which will be discussed in detail in the next section. An alternative explanation for the trend shown in Figure 4 is that the early authors writing in vernacular Chinese may indeed be more influenced by western languages and education than the later ones. Most of the early advocates and practitioners of written vernacular Chinese had studied overseas. In contrast, from 1950-1970, China was going through events like the Cultural Revolution, which greatly repelled western influence from China. It was not until 1978 when the Reform and Opening-Up Policy was implemented that Chinese people were open to western exposure again. Figure 4 does show a trend corresponding to the story just told: the Shannon indices were high before 1950, which suggests high western influence; then between 1950-1980, the indices decreased; after 1980, the indices rose again. Nevertheless, more data points are needed to justify the existence of such a trend.

ANALYSIS OF NEWSPAPER ARTICLES

News Year	Average Shannon Index	# Articles	Total Characters	Average Characters	Standard Deviation
1947-1948	0.1686	27	17642	653.4	560.8
1981	0.154	21	20658	983.7	1574.1
2000	0.1504	20	17512	875.6	1021.9
2014	0.1507	20	16015	800.75	508.4

Table 2. The weighted Shannon Indices of articles in different years from the newspaper People's Daily.

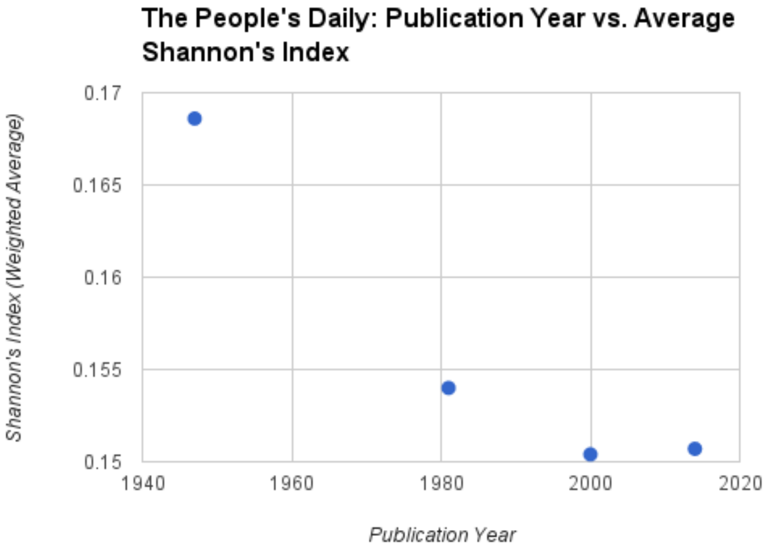


Figure 5. The average Shannon Indices plotted against publication years for random news articles from the People's Daily.

The weighted average entropies of articles from *People's Daily* are shown in Table 2 and Figure 5. In contrast to the results for novels, Figure 5 shows that the flexibility in POS usage did not increase as people are more exposed to western languages and culture. Nevertheless, more data points are needed to consolidate this conclusion.

Westernization - Potential Improvements

The Shannon Indices analyses results are not very conclusive due to several reasons. The first one is the lack of data. I was only able to perform analyses on 9 novels and 4 sets of news data of about 20,000 characters each. Collecting more data points and larger data sets will help establish more conclusive relationships between time and POS variability. Another variable to control for is the length of novels. Currently, the 9 novels chosen varied greatly in length, and it is still unclear whether length can influence the Shannon Index measures.

Another potential improvement is to make the POS variability measure more fine grained. Currently, the Shannon Index measures all kinds of transformations among various POS components. However, Chen (2005) specified a group of transformations as where English has influenced Chinese on, so it is possible that only a certain portion of the total POS variability current measured is significant for the hypothesis.

Citation

1. Wu, Jun. "Chapter 1. Word and Language vs. Digits and Information." *The Beauty of Math (*数学之美*)*. Beijing, China: Posts & Telecom, 2012. 10-11.

2. Bi, Lijun. "Chinese Language Reform and Vernacular Poetry in the Early Twentieth Century." *International Journal of Business and Social Science* 3.24 (2012): 56-65.

3. Ying, Shao, Zhangting Xie, and Yi Xiao. *Feng Su Tong Yi ()*. Taipei: Zhongguo Zi Xue Ming Zhu Ji Cheng Bian Yin Ji Jin Hui, 1978. (This print edition is published in 1978, but the book was originally written between 153-196 AD.)

4. Feng, Shengli. "On Modern Written Chinese." *Journal of Chinese Linguistics* 37 (2009): 145-61.

5. Chan, Kam Wing and Peter Bellwood. "China, Internal Migration." In Immanuel Ness (ed.). *The Encyclopedia of Global Migration*. Blackwell Publishing, 2011. 1–46.

6. Xiong, Ying. "The Influence of Dialects on the New Internet Slang Words in Modern Chinese (浅析方言对现代汉语网络新词汇的影响)." *Jiannan Literature: Reading Classics (*剑南文学*:*经典阅读*)* 2 (2011): 84.

7. Xiao, Qiang, and Perry Link. "In China's Cyberspace, Dissent Speaks Code." *The Wall Street Journal* [New York City] 4 Jan. 2013.

8. Yu, Guangzhong. "On the Westernization of Chinese (论中文之西化)." *On the Watershed (*分水岭上*)*. Taipei: Jiuge, 1981. 115-33.

9. Chen, W. "Yuyan jiechu yu yuyan bianyi: Lun Ying-Han fanyi dui Xiandai Hanyu yufa de yingxiang [Language contact and language change: On the influence of English-Chinese translation on Modern Chinese grammar]". *Liaocheng Daxue Xuebao (Shehui Kexue Ban)* 1 (2005): 85–88.

10. People's Daily [Beijing]. People's Daily (1946-Present). Web. <http://www.oriprobe.com/PeoplesDaily.shtml>.

11. Tseng, Huihsin, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. A Conditional Random Field Word Segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing* (2005).

12. Tseng, Huihsin, Daniel Jurafsky, and Christopher Manning. Morphological features help POS tagging of unknown words across language varieties (2005).

Appendix

The text files used for analyses in the “Westernization” section can be retrieved from:

https://www.dropbox.com/s/41iyz6513gv1uvr/Evol_Chinese.zip?dl=0

The source code used to analyze texts in the “Westernization” section can be found in the following GitHub Repository:

https://github.com/Lil-Sun/POS_Entropy_Chinese