愛知淑徳大学論集－文学部篇－　第 48 号　2023.3　97-106

How closely are language and culture related? －a statistical analysis of language and cultural evolution－

若　山　真　幸
WAKAYAMA Masayuki

Key Words：corpus linguistics, language and cultural changes, n-gram, collocation

1. Introduction

This paper briefly discusses interactions between language and cultural changes from the viewpoints of a corpus-driven analysis. Generally speaking, corpus linguistics examines spoken and written texts of a given language based on a collection of text data and describes grammatical characteristics and actual usage. In particular, the frequency of word n-grams has revealed hidden collocational, semantic, and even pragmatical patterns of a given language. These discoveries would not have been possible to do without the use of computational and mathematical methods.

There are various types of corpora. For example, the Corpus of Contemporary American English (COCA)[1], which is said to be the largest data set (one billion words), is a collection of present-day American English. In addition, the Corpus of Historical American English (COHA) contains historical data on American English from the 1820s to the 2010s, and the corpus of Global Web-based English (GloWbE) covers different varieties of English from 20 countries. The purposes of using corpora vary among linguists. For example, lexicographers attempt to find out the current usage of English to compile a better dictionary; some grammarians look at the differences in synonyms by looking into their collocational patterns; sociolinguists have been working on some varieties of English (i.e. regional variation or genres/styles). Linguistic data are also helpful for even schoolteachers, who can show actual and authentic expressions of English in their classes. However, few studies have been made on this topic from statistical and objective viewpoints although it has long been said that culture (more precisely, 'language culture') and language are strongly associated with each

[1] English-corpora.org (https://www.english-corpora.org)

other. This is mainly due to the discrepancy in research methodology between the two fields.

The goal of this paper is to uncover hidden and obscure relationships between language and cultural changes based on corpus and statistical analyses. The paper is structured as follows. Section 2 discusses how strongly word implications are associated with background knowledge. In section 3, I will present a study of *Culturomics* introduced by two linguists. Section 4 explores four case studies from the perspective of a corpus-driven analysis. Section 5 summarizes the discussions with concluding remarks.

2. Implications of words

Before the main discussion, we would like to mention the implications of words. It is widely accepted that word meanings can be divided into denotation and connotation. The former is a literal meaning of a given word. For example, the denotation of *white* is a color name. On the other hand, connotation refers to positive or negative associations which words invoke. For example, one of the connotations of *white* is said to be 'pure' or 'innocent'. It is important to note that connotations usually derive from cultural or historical backgrounds. In other words, semantic associations are not homogenous across languages, cultures, and generations. Therefore, it is possible to claim that cultural changes over time have an influence on meaning changes of meanings.[2]

Next, consider a more well-known example: deictic references. The definition of *a civil war* is a war between opposing groups such as a government and its people. The reference to *THE civil war* could be diverse among nations, based on their historical backgrounds. In addition, when this expression is capitalized, it refers to a specific war. In particular, American people refer to *the Civil War* as the war from 1981 to 1985, which was caused by the dispute over slavery. On the other hand, for British People, *the Civil War* refers to the Puritan Revolution, which dates back to the 15th century. Finally, Spanish People imagine their *Civil War* fought from 1936 to 1939 led by General Franco. This is how the interpretation deriving from the referential difference heavily depends on cultural and social backgrounds.

3. Language change means cultural change and vice versa

First of all, we will introduce a study conducted by Erez Lieberman and Jean-Baptiste (E&J2013), who attemped to uncover hidden connections between language and social changes based on Google Books Ngram Viewer[3]. It is widely believed that the United States of America was united into 'one' after the Civil War. However, no one can tell exactly when

[2] The story of *nice* is famous and interesting.
[3] Google Books Ngram Viewer (https://books.google.com/ngrams/)

the strong bond of the nation was established. There seems no objective evidence to prove it. Rather, it is extremely difficult to identify the time because this is a psychological phenomenon hidden in our minds. Nevertheless, the two scientists assumed that the growing number of '*the United States of America was*' is a strong linguistic sign. Note that both '*the United States of America was*' and '*the United States of America were*' were used to refer to the nation a long time ago. Over time, the singular form became more prevalent. Needless to say, a singular form is the correct form to mean the U.S.A. Based on this assumption, they investigated Google Books Ngram Viewer to demonstrate this national unification by showing the change of their frequencies in use. Their search queries are '*The United States was*' and '*The United States were*' from 1800 to 2019. Figure 1 illustrates their transitions.

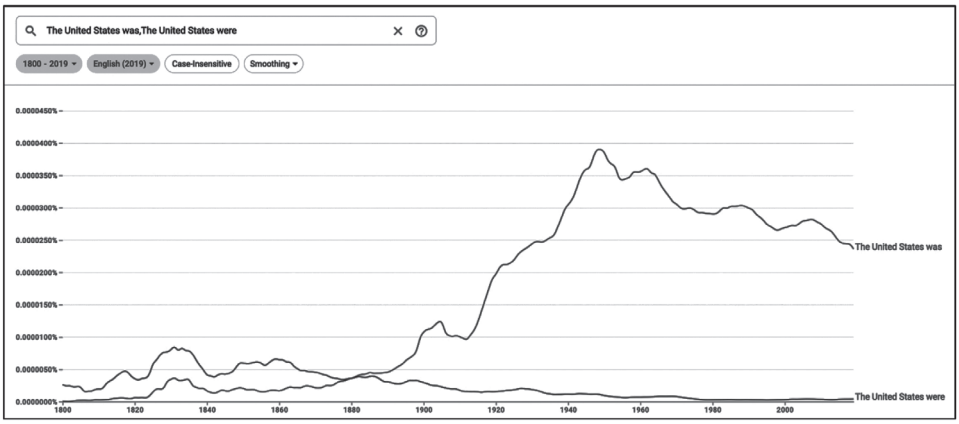

Figure 1: the trends in the frequency of '*The United States was*' vs. '*The United States were*'

At the beginning of the 1800s, the plural form was much higher, but the number of the singular form surpassed the plural one in 1880, as shown in Figure 2 below. Then, its frequency began to increase significantly after 1920. After that, the use of '*the United States was*' is increasingly dominant and the plural form is rarely observed now.

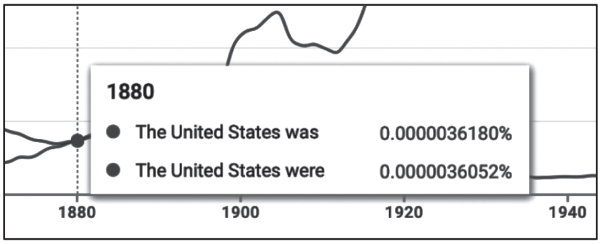

Figure 2: the year when '*The United States was*' beat the competitor.

E&J (2013) pointed out a time lag between the end of the war (in 1865) and the shift from the singular to the plural form in 1980. What does it mean? They speculate that our emotional changes progressed gradually as the political ties of the United States became stronger and stronger. Of course, we cannot say that there is a causal relationship between this language change and the change in national identity. Furthermore, this might be just speculation. Nevertheless, their study shed a light on the possibility that the statistical analysis of languages can reveal social or cultural changes.

4. Some cases

This section deals with some examples to show that language change is strongly associated with cultural change based on the investigation of three types of corpora: COCA, COHA, and Google (Books) Ngam Viewer.

4.1 Implications of *war*

Like *the Civil War*, the general term '*war*' also has different implications between American and British English, as shown in Figure 3 just below.

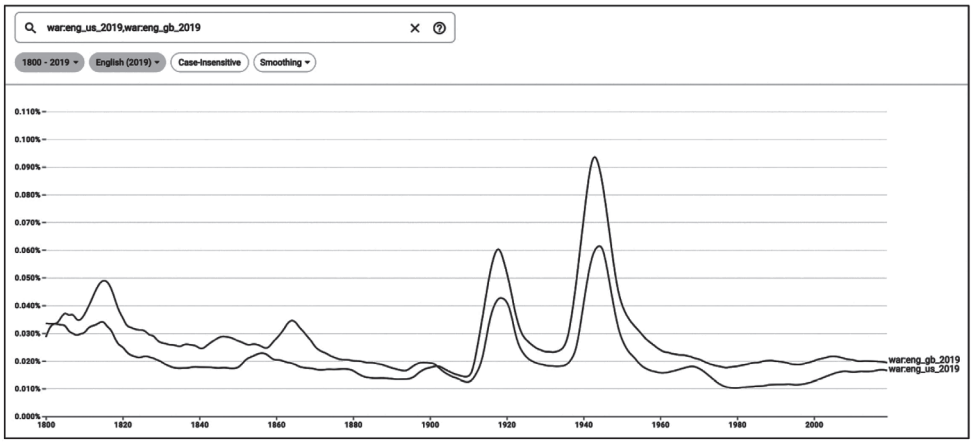

Figure 3: transition of '*war*' (1800-2019)[4]

The frequencies of certain words in different languages may vary even in the same period. For example, the peaks observed around 1920 (the period of World War I (WWI)) and 1940 (World War II (WWII)) are higher in British English than in American English. Why do the two languages yield different results for the same occurrence?

This is probably due to the different impacts and consequences of two wars on each

[4] The two lines indicate two types of data. The names 'eng_us_2019' and 'eng_gb_2019' refer to American English and British English collected by 2019, respectively.

country. That is to say, with respect to WWI, which began in 1914, the European continent was the main battlefield. Later, European countries were very busy with post-war problems. On the other hand, at the beginning of the war, the United States did not initially participate in the battles and later joined the war against Germany in 1917. This means that this '*war*' has more important implications for the United Kingdom than the United States. As a result, the frequency of *war* (i.e. times of mention) became higher in British English than in American English around 1920. Similarly, how much they were involved in WWII becomes contrasting here again. The damages by the German army were much more disastrous in England than in the United States, which did not fully enter the war until 1941. Furthermore, look at the peak during the 1860s and 1880s. This is definitely the period of the Civil War in the U.S. Therefore, the number of mentions should be higher in the US than in the UK. In this way, the meanings of a given word are strongly influenced by social background. In the following section, we will argue that a semantic change is visualized through a corpus-based analysis.

4.2 The meanings of the adjective *gay*

The meanings of the adjective *gay* are undergoing a change. Currently, the most commonly accepted definition is 'relating to sexual desire or attraction to people of the same sex' or someone who has such a sexual orientation. Historically, this word was borrowed from Old French around 1200-1300 with the meaning of '*happy*'. It is reasonable to assume, therefore, that people in ancient times used *gay* with this meaning and would become surprised to know that we use the other meaning. Interestingly, LDOCE[5] describes that the meanings such as *bright, attractive*, and *cheerful* which this adjective originally had are already old-fashioned. Based on the discussion, I made two groups, as shown in (1).

(1) Definitions of *gay*

a. if someone, especially a man, is gay, they are sexually attracted to people of the same sex

b. (old fashioned) bright, attractive, cheerful, or excited

How do we illustrate its semantic change throughout history? So far, it was difficult for even lexicographers to identify exactly when the original meaning became obsolete. (On the contrary, it is much easier to identify when the new usage emerged.) According to English-corpora.org[6], the Corpus of Historical American English (COHA) clearly reveals the

[5] *Longman Dictionary of Contemporary English*, 6th Edition

[6] This analysis comes from 'INSIGHT INTO VARIATION'. Visit the following link to see a more detailed discussion. (https://www.english-corpora.org/variation.asp)

transitions through collocations. Table 1 below is the list of the top 15 collocates that occur with *gay* within a specified range (before and after 4 words of *gay*).

First, we found from Table 1 that the sexually oriented meanings of gay became more prominent since the 1980s, because the word began to occur with *LESBIAN(S)*, *RIGHTS*, *MARRIAGE*, and *BISEXUAL*, and their occurrences are the highest in the 21st century.

On the other hand, the cooccurrence of *gay* with *BRIGHT*, *FLOWERS*, *LAUGH*, and *COLORS* had their peaks around the 1800s and went down gradually since then.

Table 1: cooccurrence of '*gay*' (All categories)

		ALL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
1	LESBIAN	260						1										1	8	70	82	98
2	GAY	260	2	5	6	7	16	7	12	9	13	2	6	22	25	6	14	8	8	18	35	39
3	RIGHTS	222																7	20	52	62	81
4	MARRIAGE	211			1		1	1					1	1		1		1	20	52	62	81
5	BRIGHT	186	5	8	10	14	13	23	12	15	12	4	12	12	15	12	9	6	3			1
6	FLOWERS	160	5	14	11	18	10	21	17	7	13	10	11	7	5	6	1	3		1		
7	LAUGH	145	3	8	5	15	13	13	14	9	12	14	9	13	4	4	5	5	3			1
8	COLORS	134	3	6	4	13	13	9	9	10	5	7	11	7	17	8	5	6	1			
9	GRAVE	132	6	15	14	10	14	8	13	13	18	9	5	4	1	1					1	
10	LESBIANS	125															1		6	43	38	37
11	LAUGHTER	92			5	5	7	6	8	4	6	15	9	11	4	2	3	2	3	1		1
12	GALLANT	91	8	11	12	4	10	8	9	6	6	1	9	4	2	1						
13	SPIRITS	86	2	3	9	8	7	7	9	8	6	6	4	9	5	2	1					
14	BISEXUAL	83																	8	10	15	50
15	BRILLIANT	80	3	9	6	11	9	8	5	3	5	4	3	3	5	5	1					

Next, consider the bigrams[7] of gay and the following noun, as shown in Table 2. In the left column (the usage from the 1820s to 1890s), as expected, this adjective was likely to occur with words whose meaning is (1b), such as *LAUGH*, *COLORS*, *VOICE*, *FLOWERS*, *TONE*, and *DRESSES*. In the right column (the usage from the 1990s to 2010s), almost all words such as *MARRIAGE*, *RIGHTS*, *COMMUNITY*, *BAR*, *PRIDE*, and *COUPLE*, mean (1a).

Table 2: the bigram list of '*gay*' (with a noun)

	1820s-1890s	1990s-2010s
1	COMPANY	MARRIAGE
2	LAUGH	RIGHTS
3	COLORS	MEN
4	WORLD	COMMUNITY
5	PARTY	BAR
6	ATTIRE	GUY
7	SPIRITS	GUYS
8	VOICE	PRIDE
9	FLOWERS	COUPLES
10	TONE	COUPLE
11	THRONG	BARS
12	SEASON	MAN
13	SCENES	SEX
14	DRESSES	SOLDIERS
15	CIRCLE	SON

[7] *gay* + any nouns

Moreover, our investigation with Google Ngram Viewer can illustrate more intriguing changes in *gay* terms. In Figure 6, we searched three expressions like *gay community, gay rights*, and *gay marriage*. Among the three, *gay community* went up faster than any other terms, and then *gay rights* went up a little bit later, and finally, *gay marriage* increased drastically around 2000.

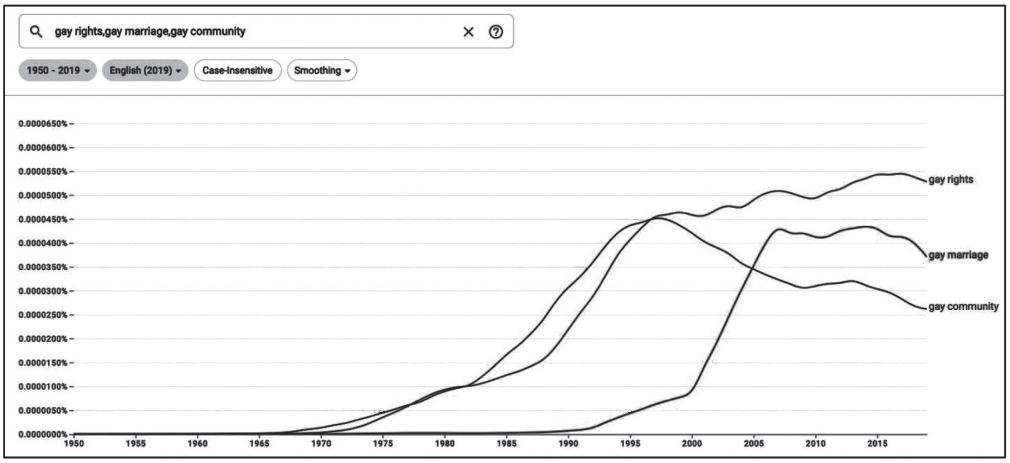

Figure 4: the transitions of *gay* idioms

This order is quite meaningful. That is to say, Figure 4 represents the progress of the acceptance of LGBT culture. People began to build a gay community and then stood up for their rights, and eventually gained one of the rights: same-sex marriage. In other words, as LGBT culture grows, new words are born and begin to be used more frequently in line with cultural development.

(2) progress of LGBT culture
community (1970) => right (since 1990) => marriage (since 2000)

4.3 quantifiers of mass nouns

Let us change the topic. Mass nouns are generally uncountable and thus cannot be used with an indefinite article *a* nor a plural ending *-s*. Instead, some quantifier is required before the nouns to indicate the amount or quantity. For example, *a piece of* is placed before *information*, and *a cup of* or *a glass of* often occurs with *water*. This is how mass nouns can be measured.

Now, look at the relationships between quantifiers and mass nouns from a different viewpoint. What noun does a given quantifier modify? For example, what noun, in particular,

does *a piece of* or *a barrel of* select? You can find the top 10 nouns easily if you use a wildcard (*) such as *a barrel of* * in Google Ngram Viewer.

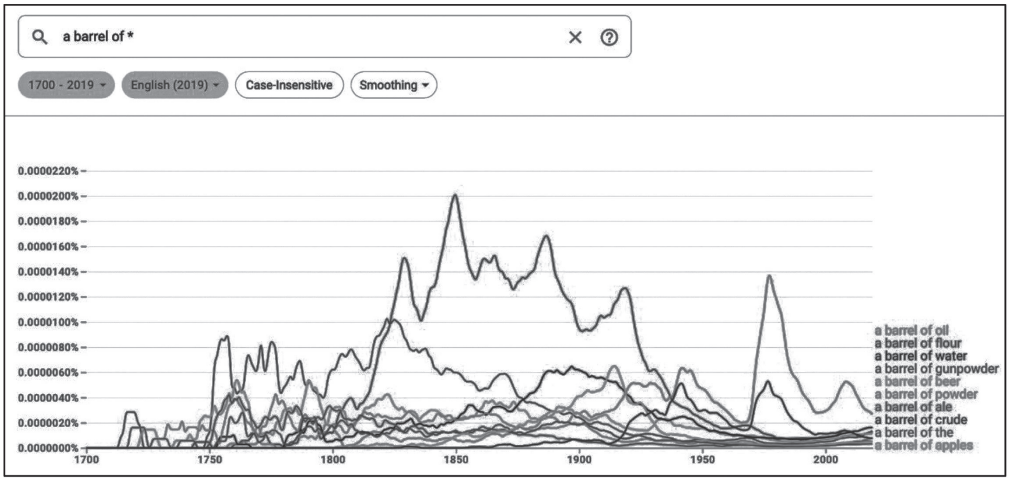

Figure 5: Transition of '*a barrel of X*' (1700-2019)

In Figure 5, it is shown that the most frequently used 4-grams in 2019 is *a barrel of oil*. In addition, *a barrel of crude* (the 8th) should be included in the same group (i.e. *crude oil*). Interestingly, their numbers were extremely low 120 years ago (the 3rd place) and rarely observed before 220 years ago. This is strongly related to the development of the oil industry. Companies in the United States established oil refineries one after another around the mid-1800s and gained commercial success. After that, oil became one of the most essential products used not only for fuel but plastic commodities. In addition, oil sometimes had significant impacts; the highest peak is seen around 1975-1980 when the 1973 oil crisis occurred. In this way, types of nouns modified by *a barrel of* will vary in each historical period.

Incidentally, what words occurred with the quantifier in the past? *flour* was most frequently used from 1800 to 1920, but *gunpowder* was the most common before 1800. Our investigation was able to illustrate such changes in shipping goods throughout history. At school, we found the matter of mass nouns and quantifiers a boring subject. Nonetheless, the changes in their collocations demonstrate interesting transitions.

4.4 The change of *X-free* expressions

We need a new word when a new trend begins. In this section, we will examine our health-conscious life and language changes with a focus on the development of *X-free* words.

Although we are always careful about our health, we are at risk of health problems in our daily life. For example, in English-speaking countries, obesity, overweight, stroke, alcohol abuse, and tobacco are common health issues. Therefore, people like to buy fat-free foods and sugar-free snacks. In these examples, words such as *X-free* denote *something free from X* or *something without X*.

Now, let us see their changes since 1900. We picked up seven *X-free* words relevant to health problems[8].

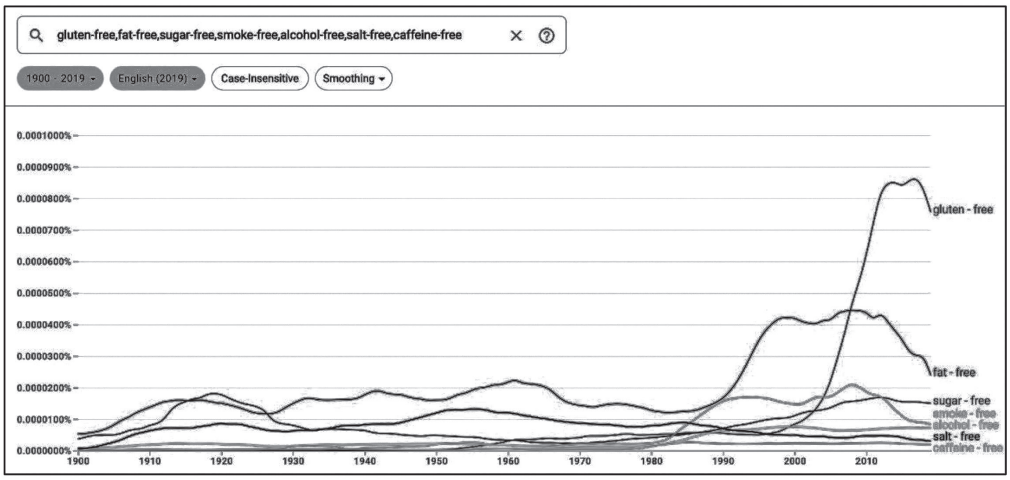

Figure 6: the transition of *X-free* expressions (1900-2019)

As expected, the term *gluten-free* is the currently most popular expression. According to Figure 6, this vocabulary emerged in the 1950s and extremely increased in this century[9]. This trend is consistent with the rise of a gluten-free diet in the U.S.A. Furthermore, we also found that people had been particularly concerned about *fat-free*, *sugar-free*, and *salt-free*, which were very common until the surge of *gluten-free*. On the other hand, *alcohol-free*, and *smoke-free* are not as frequent as we expected. With respect to the former, people do not purchase alcohol-free products because they pay attention to alcohol abuse or simply stop drinking to recover from alcoholism. Thus, the frequency of *alcohol-free* is constantly not high throughout history. As for *smoke-free*, the number increased around the 1980s. This is quite consistent with the tobacco control movement in the U.S.A. People became aware of the link between smoking and lung cancer, as well as the harmful effects of secondhand smoke, which led to the anti-smoking movement.

[8] Seven words are *gluten-free, fat-free, sugar-free, smoke-free, alcohol-free, salt-free, and caffeine-free.*
[9] According to the COHA, this expression was first attested in an article of *New York Times* in 1983.

5. Concluding Remarks

The goal of this paper was to uncover hidden and obscure connections between language and cultural changes through corpus and statistical analyses. The present study provided four instances. First, Google Ngram Viewer demonstrated different implications of a given word. Second, the present study clearly showed that collocation changes reflected semantic changes. Next, it turned out that adjectival modification of quantifiers with mass nouns was strongly influenced by some historical events. Finally, the growth of *X-free* expressions was synchronized with health trends. In this way, a statistical analysis of languages, using corpus and n-gram techniques, can shed new light on the interactions between language and cultural changes.

References

Erez Lieberman, Aiden, and Jean-Baptiste, Michael (2013) *Uncharted: Big Data as a Lens on Human Culture*, Riverhead Books: New York.

Longman Dictionary of Contemporary English, 6th Edition (2014) Pearson Education: London.

Sources

English-corpora.org

https://www.english-corpora.org/corpora.asp

The Corpus of Contemporary American English (COCA)

https://www.english-corpora.org/coca/

The Corpus of Historical American English (COHA)

https://www.english-corpora.org/coha/

The corpus of Global Web-based English (GloWbE)

https://www.english-corpora.org/glowbe/

The Google Books Ngram Viewer

https://books.google.com/ngrams/

(All accessed on December 3 and 4, 2022)