Google N-grams Viewer and Food Idioms

Sarah V. C. Ribeiro¹ and Paula L. C. Lima²

¹ Instituto Federal do Ceará (IFCE) and Universidade Estadual do Ceará (UECE), Fortaleza-CE, Brazil – sarah.virginia@aluno.uece.br

² Universidade Estadual do Ceará (UECE), Fortaleza-CE, Brazil – paula.lenz@uece.br

Abstract. The purpose of this study is to use the Google Books N-gram Viewer as a tool of investigation in order to show the frequency of use and possible obsolescence of 16 food idioms in English. We analysed the percentage of use for the first and latest records and tallest spike of each idiom, as well as the period of time they occurred. We found evidence that some of them are in very little use and with a frequency of use in decrease, while others follow the opposite direction. We also compared these results with Webcorp occurrences of the same idioms and the findings were similar for most of them. The Google N-gram Viewer was found to be an appropriate tool to analyse the frequency of use of idioms.

Keywords: Frequency of Use, Obsolescence, *Corpus* Linguistics.

1 A first view

Idioms are part of our every-day language and, as such, they are an important topic that relates to different fields of study, such as machine translation, lexicography and second language acquisition, among others. They belong to figurative language and, for a long time, were traditionally considered as frozen constructions, but “new theories on metaphor comprehension have shed lights upon idiom studies, encouraging different perspectives [6]. These gave more emphasis to their cognitive essence rather than their semantic origins. Scholars such as Lakoff, Gibbs and Giora, among others, have brought important insights on the mechanisms of idiom comprehension. The number of studies on idioms has constantly increased in the last 5 decades [6]. Aspects like transparency, decomposability, salience and conventionality play an important role in order to determine idiom comprehension. Familiarity is another aspect, which is directly related to the frequency of use of this type of language.

Despite their importance, it is sometimes hard to know whether some idioms are still in use or have become obsolete. Many times, they are only seen in dictionaries, as a record of an expression that was highly used for some time, but has somehow fallen out of interest. The purpose of this study is to investigate the appropriateness of using the Google Books N-gram Viewer (GBNV, hereafter) to verify the frequency of use and possible obsolescence of 16 English idioms that have food names in their composition. This computer tool has more than 5 million books published from 1500 to 2008, contains 500 billion words from various monograph/book materials found in the Google Books collection as its *corpora*, and shows the occurrence of words (n-grams) or short phrases (up to 5 words) in the form of a plotted line chart.

2 A better view

Since GBNV's first release in 2009, many of its positive and negative aspects have been discussed. Some of the negative critics concerned the quality of the optical character recognition (OCR) software and other conditions that reduced digital image quality [8], or the overabundance of scientific literature, or yet, the messy metadata [9]. One of the positive aspects was the size of the *corpora* compared to other *corpora* available at that time. Although some scholars were excited about the possibilities of such large *corpora*, several others were sceptical about its dependability [3]. Another positive aspect was that Google gave the possibility of freely downloading the raw data available. According to Davies [4], one thing the GBNV 2009 version did well was "to show the frequency of a given word or exact phrase over time, which provides insight into lexical shifts in the language". Cohen [3] states that the best possibilities of using GBNV might be for longer grams "since they begin to provide some context." Their vision endorse the appropriateness of using this tool to achieve the objective of this study.

The GBNV 2012 release brought advances, such as the improvement of the OCR system and the inclusion of wildcards and other features, bringing more functionality to the searches [5]. Thus, this is the version we used for this analysis.

The 16 idioms analyzed here are licensed by the DIFFICULTY/EASINESS IS A FOOD DIFFICULT/EASY TO HANDLE/DIGEST metaphor, and were among those taken from two dictionaries of idioms [1][2] which make up the *corpus* of our broader study on food-idiom machine translation and conceptual metaphors. The obsolescence or frequency of use of the idioms may influence the quality of human or machine translation, more so for machine translators that use statistical paradigms.

For this study, we used the GBNV filters: time span from 1800 to 2008, with 0 smoothing, and with the case insensitive box activated (although it was not always possible to use this function, e.g. with wildcards, a limitation of the tool itself).

We searched each idiom individually and, for a few, we searched more than once since there was the possibility of different spellings (e.g. **sell like hot cakes/hotcakes**) and other variations (e.g. **get/got out of a jam**). All the graphs were analyzed and their percentages taken notes. We checked all the sentences (books) given for each idiom to confirm their idiomatic use. In order to validate the results, we crossed them with the number of occurrences of the same 16 idioms generated by the Webcorp whose idiomatic use we have previously confirmed. The Webcorp [7] is an online search engine, which allows access to the World Wide Web as a *corpus*, making it possible to extract concordances of the word(s) searched and generating much updated results.

3 A detailed view

An example of the charts plotted for the searches is presented in Fig. 1, the idiom **not cut the mustard anymore**. As shown, its first record occurred in 1968, its tallest spike was in 1981, with a percentage of use of 0.000001200%¹. The chart also shows other spikes during the period of use searched, and the percentage of use of 0% in 2008, which indicates that this idiom might be obsolete.

¹ The frequency is calculated according to the number of words in the GBNV *corpus*.

Fig. 1. Chart plotted for the idiom **not cut the mustard anymore**.

It is important to mention that GBNV only considers n-grams that occur in, at least, 40 books; otherwise, it plots a flat line [5]. Table 1 shows the percentage charted for the first and latest records (2008 for all), and the tallest spike (the highest percentage of frequency). It also brings the number of occurrences generated by the Webcorp.

Table 1. Frequency of use percentages of first and latest record, tallest spike in the GBNV and number of occurrences in the Webcorp²

Food idiom/Expressions ³	First record (%)	Tallest Spike (%)	Latest record (%)	Webcorp
sell like hotcakes/hot cakes	0.0000005996	0.0000023032	0.0000009362	83
walk on eggs/eggshells	0.0000005732	0.0000043996	0.0000016117	69
upset the apple-cart/apple cart	0.0000006810	0.0000044341	0.0000017583	60
a/no piece of cake	0.0000015499	0.0000296017	0.0000212620	58
a hard nut to crack	0.0000006048	0.0000062144	0.0000026118	44
a (pretty) kettle of fish	0.0000015515	0.0000086885	0.0000022535	43
a cake-eater/cake eater	0.0000002858	0.0000010682	0.0000000257	40
get out of a jam	0.0000003597	0.0000003996	0.0000002390	29
handle the hot potato	0.0000004702	0.0000004702	0.0000000216	15
not cut the mustard anymore	<u>0.0000000278</u>	<u>0.0000001200</u>	<u>0.0000000000</u>	14
butterfingers	0.0000002028	0.0000016333	0.0000004055	9
have a hot potato	0.0000001782	0.0000001941	0.0000000108	<u>3</u>

The analysis of the data revealed that, from the 16 idioms searched, 3 did not show any results (**left with * hot potato; have a lemon on your hands; and give * the/a hot potato**), a result similar to the number of occurrences generated by the Webcorp (4; 4; and 1, respectively). That does not necessarily mean they were not used at all, but that their frequency of use may have been lower than the 40 records necessary to be charted

² The highest results are in **bold**, and the lowest, underlined.

³ The frequency of use percentage from the GBNV includes non-idiomatic expressions.

by the tool. Nevertheless, the lower frequency can be a sign that these idioms are on the process of becoming obsolete. One of the idioms analysed, **a small beer**, showed a high frequency of use, but, after checking the sentences in which it appeared, we noticed that its use was not idiomatic in any (*e.g.* a small beer garden), so it was not included in the table, along with the 3 others that generated no results, afore mentioned.

The highest first record percentage found was for the idiom **a (pretty) kettle of fish**. The lowest first record was for the idiom **not cut the mustard anymore**. This idiom also had the lowest latest record and lowest tallest spike. The highest latest record and tallest spike were, by far, for **a piece of cake**, but that included a large percentage of non-idiomatic sentences (31.7%). On the other hand, **no piece of cake** had only 8% of non-idiomatic use. Some idioms had all, or nearly all, of the sentences in which they appeared with idiomatic use. A possible explanation for that may be the level of idiomaticity. In total, we analysed 1,517 sentences/books. From these, 75.4% were idiomatic, 22.3% were non-idiomatic, and 2.2% could not be accessed. The results from GBNV, for both the highest and lowest percentages, seem to be corroborated by the number of occurrences generated by the Webcorp, taking into consideration that these include only the occurrences where we identified idiomatic use. Although we can identify the (non) idiomatic use of each idiom, we cannot subtract it from the graphs.

Concerning the years, the idiom with the oldest first record was **a pretty kettle of fish** (1806). The one with the most recent first record was **not cut the mustard anymore** (1968). **Walk on eggshells** was the idiom with the most recent tallest spike (2007), so still probably highly used; while **walk on eggs** had its tallest spike much earlier (1843). The idiom with the oldest tallest spike was **a kettle of fish** (1824). The idioms whose frequency of use was falling in 2008 were **sell like hotcakes**, **a/no piece of cake**, **walk on eggshells**, **get out of a jam**, **butterfingers**, **a cake-eater/cake eater** and **upset the apple cart**. The idioms that showed a tendency to rise in frequency in 2008 were **sell like hot cakes**, **walk on eggs**, **have a hot potato**, **a hard nut to crack**, **a (pretty) kettle of fish**, **handle the hot potato** and **upset the apple-cart**.

4 A final view

The data from the GBNV show results similar to those generated by the Webcorp, as far as the frequency of use of the 16 idioms is concerned: 3 idioms did not show any results, 1 showed only non-idiomatic results, and 1 had 0% of frequency in 2008, showing that they might be obsolete or in the process of becoming so. The other idioms presented different percentages of use, half of them plotted a decrease of use in the last years, and half plotted an increase.

Similarly to the Webcorp, the limitations concerning the use of the GBNV are that the results include sentences where the n-grams searched are not used idiomatically, making it necessary to check each sentence/book; and many of the examples come from dictionaries - therefore not necessarily an example of the idiom in use, but its explanation. In addition, if idioms are larger than 5 words, the search can become more complex. Nevertheless, GBNV was found to be an appropriate tool to analyse the frequency of use of idioms and to identify a possible process of obsolescence.

References

1. Camargo, S. e Steinberg, M. Dictionary of Metaphoric Idioms English-Portuguese. E.P.U., (1990).
2. _____. Dicionário de Expressões Idiomáticas Metafóricas Português-Inglês. E.P.U., (1989).
3. Cohen, D.: Initial thoughts on the Google Books N-gram Viewer and datasets. <http://www.dancohen.org/2010/12/19/initial-thoughts-on-the-google-books-ngram-viewer-and-datasets/#comment-67769>, last accessed on 2017/06/10.
4. Davies, M.: Making Google Books n-grams useful for a wide range of research on language change. John Benjamins Publishing Company (2014).
5. Google Books Homepage, <https://books.google.com/ngrams/info>, last accessed on 2017/06/15.
6. Liu, W. and Shen, H.: CiteSpace II: Idiom Studies Development Trends. Journal of Arts and Humanities (JAH), 2 (2), March, 85-97 (2013).
7. Webcorp Homepage, <http://www.webcorp.org.uk/live>, last accessed 2017/08/10.
8. Weiss, A.: Google Ngram Viewer. California State University, Northridge (2015). <http://www.scholarworks.csun.edu/bitstream/handle/10211.3/173281/NGram-v2-1.pdf;sequence=1>, last accessed on 2017/06/04.
9. Zhang, S.: The pitfalls of using Google ngram to study language. <https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram/>, last accessed on 2017/05/29.