BEN-GURION UNIVERSITY OF THE NEGEV
FACULTY OF ENGINEERING SCIENCES
DEPARTMENT OF INFORMATION SYSTEMS ENGINEERING

# Temporal Trends in the Use of Multi-word Expressions

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE M.Sc DEGREE

By: Tal Daniel

6 July 2015

BEN-GURION UNIVERSITY OF THE NEGEV
FACULTY OF ENGINEERING SCIENCES
DEPARTMENT OF INFORMATION SYSTEMS ENGINEERING

# Temporal Trends in the Use of Multi-word Expressions

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE M.Sc DEGREE

By: Tal Daniel

Supervised by: Prof. Mark Last

Author: …….. Date: ……..

Supervisor: …… Date: ……..

Supervisor: …… Date: ……..

Chairman of Graduate Studies Committee: …… Date: ……

6 July 2015

## Abstract

When we speak our native language, we naturally use current language terms and phrases, and rarely use outdated expressions. The task of differentiating between outdated expressions and current expressions is not a trivial task for foreign language learners, and could also be beneficial for lexicographers, as they look for new expressions.

Thus, we are interested to distinguish between outdated expressions and their current synonyms within a given lexicographic corpus. We focus on Multi-word Expressions (MWEs) of 2–3 words in English, which we simply refer to as expressions. We also try to characterize a few MWE properties, such as minimum frequency and sparsity.

Assuming that the usage (or functionality) of expressions over time can be represented by a time-series of their periodic frequencies over a large lexicographic corpus, we test the hypothesis that there exists an old–new relationship between some expression time-series – a hint that synonym expressions replace one another. In addition, we test the hypothesis that synonym expressions can be found by identifying expressions with opposing usage trends. Another hypothesis we test is whether MWEs can be simply characterized by a certain sparsity & frequency thresholds.

In order to find trends in MWE usage, we searched for MWEs that match a ready-made list of 65,450 MWEs (based on WordNet 3.0), within the Corpus of Late Modern English Texts (CLMET 3.0), which contains 333 books published between the years 1710–1920, and within the Google Syntactic Ngrams dataset, which is based on 1 Million books, published between the years 1502–2008. We detected a trend with Kendall's $\tau$ nonparametric correlation coefficient test and Daniels test for trend, which uses Spearman's $\rho$ rank correlation coefficient.

We tried to find synonym expressions by matching pairs of expressions with medium-high negative correlation between their trends, but, as this method performed poorly, we manually looked for synonyms of the 30 expressions that had the most positive trend, and synonyms of the 30 expressions that had the most negative trend, using a historical thesaurus. Then, we used the synonyms we found in order to show that old–new relationship exists for some expression pairs.

From the collected data, frequency & sparsity thresholds for MWEs were calculated. These thresholds were then used in order to find candidate expressions since 1904. The results were qualitatively evaluated.

Pairing expressions with negative correlating trends as a method to find synonym expressions performed poorly: We found only a few near-synonyms, by manual examination of expression pairs.

Using a historical thesaurus, we found some synonym expressions to expressions in our data, and this allowed us to visualize opposing trends, or interaction, between synonym expressions, as hypothesized.

We found that our ready-made list of MWEs contained expressions that were mentioned at least 0.122 times per million words, over at least 28 years in a row (1,360 occurrences out of 1.11432E+11 words in the Google Syntactic-Ngrams dataset). This may suggest a minimum period and usage (functionality) level that linguists and lexicographers test an expression against, before it qualifies as a suitable entry in a dictionary or lexicon.

## Keywords

Multi-word expressions, phrases, synonym expressions, trend detection, language change, computational linguistics, historical linguistics.

## Thanks

I would like to thank Mark Last, for his confident & patient instruction, Hendrik De Smet, for information and assistance in using the CLMET 3.0 corpus, and Mark Finlayson for his assistance in using the jMWE library. I would also like to thank Marina, my life partner, for allowing me to play with code & words, and last but not least, I thank my parents, for nourishing my curiosity.

# Table of Contents

# List of Figures

# List of Tables

# 1. Goals & Objectives

In this work we explore Multi-word Expressions (MWEs, or expressions) usage over a period of a few hundred years. Specifically, we look for English expressions with long-term decreasing or increasing usage trends, and focus on expressions of 2–3 words that exist in a ready-made list of MWEs. By finding expressions with statistically significant trend, we hope to identify a subset of expressions that have an interesting usage relationship with their near-synonym expressions – replacing them, or being replaced by them over time.

Another objective of this work is to find potential new interesting candidate MWEs in a dataset of collocations, by finding expressions not in the ready-made list of MWEs that withstand certain thresholds and have a statistically significant trend over the years.

## *1.1. Contribution*

By finding trends in expressions, we purpose & test two new hypotheses: (1) Whether synonym expressions could also be found by finding pairs of expressions with opposing trends, and (2) whether MWEs could be characterized by two simple thresholds, overlooked, or not characterized explicitly, by MWE extraction methods (time-series sparsity & normalized frequency). The two properties resemble the properties used in a TF-IDF formula (Term Frequency and Inverse Document Frequency), though we look at the sparsity, instead of document frequency. In addition, we do not combine the two values into a single formula, but use them as two separate thresholds. These values portray the minimum mention number of an MWE, and more importantly, the minimum period of time an expression is used before it is accepted as a “valid” MWE.

The latter value could be extended to differentiate MWEs from temporary multi-word terms or keywords.

This work contributes also by highlighting various expressions that have an increasing or decreasing usage trend in the English language. The expressions were revealed using the methods we developed, and each expression with a trend may be further expanded for further linguistics work. The list of expressions with a trend, along with all their related data, would be published for public use.

# 2. Related Work

## *2.1. On language research*

Since this work is concerned with examining the trends of synonymous multi-word expressions over a long period of time, it is related to the fields of natural language processing (NLP) research, computational linguistics, and historical linguistics. Historical linguistics tries to find links between words & phrases in current language to historic "snapshot" of the language. Language change is another research topic, not covered in this work, which tries to go a step beyond finding changes, and model the cultural factors that brought to these changes (Aitchison, 1991).

The field of corpus linguistics tries to build theories based on corpora analysis. The corpus of text is assumed to be representative, balanced sample of the language, so generalization would be valid (McEnery, Xiao, & Tono, 2006). Balance is reached in general-purpose corpora by “covering a wide variety of frequent and important text categories [e.g., prose, news] that are proportionally samples from the target population [i.e., the language under research]” (McEnery, Xiao, & Tono, 2006, p. 21). However, even when a text corpus is balanced, there is no objective method that can really

measure if the corpus is indeed representative of the language – the population (McEnery, Xiao, & Tono, 2006).

Corpus linguistics mainly use descriptive statistics, without performing statistical tests for significance (McEnery & Hardy, 2012). Criticism exists when using significance tests in language, since langue is not random (as many statistical tests assume), but rather follows a Zipfian distribution, and since the sample size is enormous, which affects the power of a statistical test. Therefore, caution should be used when hypothesizing about language phenomena. (Kilgarriff, 2005)

## *2.2. Multi-word Expressions (MWEs)*

### 2.2.1. Definition

Languages contain multi-word expressions (MWEs) that are compounded from a few words (lexemes). MWEs contain various types of expressions such as transparent collocations, fixed phrases, similes, catch phrases, proverbs, quotations, greetings, & phatic phrases (Atkins & Rundell, 2008). They are also used "to enhance fluency and understandability, or mark the register/genre of language use […]. For example, MWEs can make language more or less informal/colloquial (c.f. *London Underground* vs. *Tube*, and *piss off* vs. *annoy*)" (Baldwin & Nam, 2010). Heid (2008) mentions that classifying an expression as MWE involves examining the expression by "[…] formal aspects (e.g. binomials) and semantic aspects (e.g. in idioms), but also the category […] (e.g. in multi-word function words), its pragmatic use and relevance (e.g. in stereotyped comparison or proverbs, quotes and sayings) […]" (Heid, 2008, p. 340).

MWEs differ regionally/nationally, and often, cross-lingual variability makes the task of machine translation even harder, due to the use of different MWEs in other languages (Baldwin & Nam, 2010). In fact, automatic translation systems contain special dictionaries of MWE lists (Ramisch, 2013, pp. 3–4). This problem is even harder than it sounds, since there is evidence that “so-called 'fixed-phrases' are not in fact fixed” (Sinclair 1996, as mentioned by Philip, 2008). Philip (2008) emphasized that MWEs have variants, or “exploitations” that go beyond replacing a word with its semantic equivalent.

Besides dictionaries and machine translation, MWEs are important for improving the quality of OCR applications, word-sense disambiguation, part-of-speech tagging, and information retrieval.

Contrary to initial gut feeling, MWEs are quite common, as Firth hints: “MWE are habitual recurrent word combinations of everyday language” (Firth, 1957; as mentioned in Ramisch, 2013). Some expressions are even more common than single words, and they are considered relevant when learning a second language (McCarthy & Carter, 2002). The English language contains around 40% expressions (Fellbaum, 1998, as mentioned in Tsvetkov & Wintner, 2011; Jackendoff, 1997). Sinclair (1991, as mentioned in McCarthy & Carter, 2002) introduced the idiom principle, and open choice principle, which, together, allow one to choose ready-made structures and use syntax to connect these structures in sentences. These principles, according to Sinclair, have greater influence than Chomsky's rules (Chomsky, 1965, as mentioned in McCarty & Carter, 2002).

Recent research enhanced Firth's (1957) definition of MWEs. A somewhat more formal definition may be “Multiword expressions […] are lexical items that:

(a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity" (Sag, Baldwin, Bond, Copestake, & Flickinger, 2002, as mentioned in Baldwin & Nam, 2010). The idiomatic nature of MWEs "[…] refers to markedness or deviation from the basic properties of the component lexemes, and applies at the lexical, syntactic, semantic, pragmatic, and/or statistical levels" (Baldwin & Nam, 2010).

As we understand MWEs, they have common ground with *collocations*, *idioms*, and *multi-word terms*: Collocations are used as the broadest set of recurrent chunk of texts: "[…] any statistically significant co-occurrence, including all forms of MWE" (Sag et al., 2002, as mentioned in Ramisch, 2013). Note that the definition of collocations may lead to confusion, since computational linguistics research defines collocations as a subtype of MWEs (Heid, 2008). *Idioms* are another subset of collocations. Another distinction Ramisch makes is between MWEs and MWTs (multi-word terms), which are a special case of terminology: Terms can be single-word or multi-word, and they are usually present in a more technical/scientific texts (Ramisch, 2013, p. 24). As far as we understand, we can place MWEs in word-space as follows:

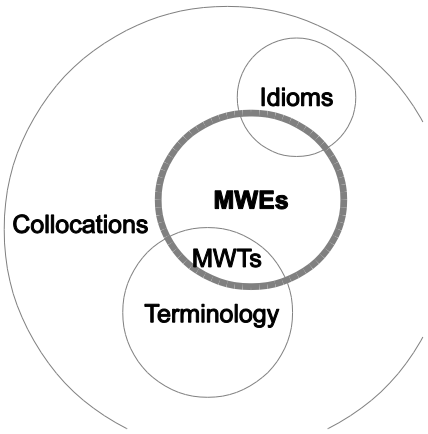


*Figure 1: MWEs location within word-space; a subset of collocations, while its borders with idioms and multi-word terms are not rigid. Inspired by Ramisch (2013).*

Since MWEs are a mixed set with multiple phenomena, we adopt the broad and practical definition that Ramisch used, based on Calzolari et al.: "[…] *phenomena [that] can be described as a sequence of words that act as a single unit at some level of linguistic analysis*" (Calzolari et al., 2002, as mentioned in Ramisch, 2013). This definition emphasizes that MWEs are a single unit, which is especially important for translation sake, as Ramisch hints. We can narrow the definition even further, though, if we consider MWEs as units that have single word synonym, or translation to another language.

Ramisch (2013) mentions some characteristics of MWEs: Some of them are idiomatic expressions (e.g. *pull one's leg*), while others "[…] have the singularity of breaching general language rules" (Ramisch, 2013, p. 2), such as *from now on*, *from time to time*, etc. They may be common names, e.g., *master key*, *vacuum cleaner*, and "sometimes the words in the expressions are collapsed and form a single word […]" (Ramisch, 2013, p. 2), like *honeymoon*, *firearm*. In addition, some MWEs has no single word replacement.

### 2.2.2. Finding MWEs

Several methods exist for finding, or extracting, MWEs from corpora. Often, researchers focus on a single kind of expressions, and length, e.g., Noun-Noun expressions of length two, in Al-Haj & Wintner (2010), or Verb-Noun idiom construction in Fazly, Cook & Stevenson (2009). Focusing on a certain kind of expressions can be achieved by crafting a tailored-characterization of these MWEs, and creating a learning system, in order to build a learning model that finally finds more MWEs within the corpus, based on the learned model. For example, Tsvetkov & Wintner (2011) suggested a method for any kind of MWEs, by training a system to learn a Bayesian model, based

on characteristics (features), such as the number of contexts the expression occurs in, its semantic variability (how flexible it is to synonym word replacements), syntactic variability, or whether a translation of the expression appears in another language.

## 2.3. Trend Detection

As languages change over time, finding trends in language is a task that resembles counting animals in nature: By watching and counting we can only estimate how many specimens exist in nature. In the same manner, every text corpus is a sample, showing a partial image of language use.

In Hebrew, there are synonym expressions, where some of them are current & prevalent than others, which were used in past centuries. We presume that the same phenomena occurs in English too. As new expressions become less, or more, frequently used, we can try to track these changes over the years by finding trends.

Identifying a trend involves a few tasks, though: One has to identify a statistically significant change in the data over time, to estimate the effect size of that change, while trying to pinpoint the exact time periods of these changes. In addition, prediction is also a possible requirement for trend analysis (Gray, 2007).

Buerki (2013) compared three methods for finding trends in a corpus. He found that Chi-square was the most flexible, had an arbitrary cut-off frequency value when stating a statistically significant change in frequency, and could alert of a trend when it occurred in some periods, compared to other methods – not only to a continuous linear increase/decrease. Chi-square outperformed, as Buerki states, other methods as coefficient of difference (D) – the sum of squares of frequencies for each period (Belica, 1996), or coefficient of variance (CV) used by Baker (2011, as mentioned in

Buerki, 2013), which ranks the terms and use an arbitrary cut-off point (e.g., first third of the ranked list). Buerki divided the corpus into only 5 periods (data points), and tested only MWEs that appeared at least four times over two periods; the frequency number was set to three times the number of documents an MWE appeared in.

Kendal's τ statistic is used as a nonparametric correlation coefficient; when the assumption of normal distribution could not be assumed, or when the correlation between two variables is non-linear. The statistic is calculated as the sum of differences between the number of data-points that are greater than a certain data-point, to the number of data-points that are smaller than it. Kendall's τ allows one can test for a statistically significant trend, but without trend positioning – stating in which period the trend occurs (Gray, 2007).

Page (1954) introduced Cumulative Sum method for change detection, which has a few variants. It is based on creating a sum of derivatives along the time axis, creating a cumulative sum line. If the data that has no statistical significant change somewhere along the time-series, the line will resemble a linear line. Alippi & Roveri (2006) compared CUSUM to the Mann-Kendall test, and found that when a change occurs, the CUSUM method finds it more quickly, and with less mistakes (either False-Positives or False-Negatives), but their experiment included an abrupt change along the time-series, which is probably not the way phenomena in language behave.

## *2.4. Synonymy*

Synonym expressions extend synonym words, where a synonym expression can replace other expression to convey the same meaning. This claim is not accurate, though, since synonyms are not perfectly similar to each other: "Synonymy, or more

precisely near-synonymy, is the study of semantic relations between lexemes or constructions that possess a similar usage" (Glynn, 2010, p. 2). While Glynn's Cognitive Linguistics research investigated differences between *annoy*, *bother*, and *hassle*, Kalla (2006) researched differences between three Hebrew words that mean *a friend*: ידיד, רע, עמית.

Uzi Ornan's book (1995) lists Hebrew words that are not used nowadays, and wrote their current synonyms, where available. His methods were not computational, though.

We focus on finding synonymous expressions with opposing trends, which is different from existing methods: Synonymous expressions can be found by using a dictionary to replace words in existing expressions with their synonymous words. Then, the resulting expression variations are searched in the text (Juska-Bacher & Mahlow, 2012).

Meusel, Niepert, Eckert, & Stuckenschmidt (2010) created a system that expands a thesaurus, by finding new, candidate words for the thesaurus, from search engine results, and using a trained classifier to verdict if there is synonymy to a thesaurus entry. They trained the classifier on 100 "new" words and their 100 candidate synonym pairs, from the thesaurus (10,000 pairs, overall). For each pair, patterns of synonymy that were found in a search engine composed the pair's feature vector. The final classifier used the top 70 patterns that implied synonymy. Though a decision tree had a precision of 98%, it was not good at separating hyponyms from synonyms. Their system was used for single words, though.

Another method to find synonym relationship is to look at the surrounding words of the term, and find another term with similar surrounding words. This method is also titled distributional similarity. Heylen, Peirsman, Geeraerts, & Speelman (2008) compared three models of this method, and found that the model that took into account the

dependency of the surrounding words to the term performed better than models that used only the bag of words surrounding the term, without dependency. The syntactic dependency model (based on Lin, 1998), performed even better for terms that had low frequency in their Dutch corpus.

Collier, Pacey & Renouf (1998) developed software tools to find synonym words in a corpus, including expressions. Synonym relationships were found by using distributional similarity, for 2-gram expressions, and then used collocation and clustering to further restrict the semantic relationship between potential synonym expressions. Though they also found expressions that had abrupt change in frequency in a certain year, they did not expand it to track trends over periods.

Mahlow & Juska-Bacher (2011) created a German diachronic dictionary by finding variations of pre-selected expressions. Expression variations were found by using patterns and by assigning expressions to types (categories). Juska-Bacher & Mahlow (2012) elaborate more on their semi-automatic method to find structural and semantic changes in German phrasemes (idiomatic MWEs): First, they found candidate phrasemes by looking at nouns with at least %2 frequency, as well as other indicators. Then, they chose select phrasemes, after manually looking into old and contemporary dictionaries. These phrasemes were found in various corpora and manually analyzed for changes. Above all, their work emphasizes the importance of manual examination, in addition to corpus-based approaches: “Fully automatic detection of phrasemes is not as yet possible, which is why lexicographers have to manually determine idiomaticity” (Rothkegel 2007, as mentioned in Juska-Bacher & Mahlow, 2012).

Liebeskin, Dagan & Schler (2013) used a semi-automatic iterative & interactive approach for creating a diachronic Hebrew thesaurus. They tried to automatically find

synonym terms for a given list of terms by using second-order distributional similarity. Then they let a lexicographer to either select synonyms, or mark terms for query expansion, which set the system into another synonym look-up iteration, based on these marked terms.

# 3. Research methods

In the following section we describe the methods we used to find MWEs using a ready-made MWE list (Kulkarni & Finlayson, 2011) and a corpus or dataset. Then, we describe how to prepare a histogram, or time-series, in order to statistically test for usage trends, and the statistical tests for finding a trend.

We detail two methods we used for finding synonyms, one is automatic, and the other manual, by looking at a thesaurus. We also describe a [simple] method for finding *candidate expressions*.

## *3.1. Finding MWEs*

Since MWE extraction is a complicated task on its own, we looked for a ready-made list of MWEs and a tool that could help process texts. We found the jMWE library suits that purpose (Kulkarni & Finlayson, 2011): It has a simple MWE detector that searches in a text, or sentences, instances of an expression that exists in its MWE list.

The MWE list contains over 65,450 MWEs, and is based on WordNet 3.0 Semcor 1.6 (updated in May 7, 2011). The expressions are not marked as historical expressions or not.

Another hypothesis we wanted to check was whether MWEs could characterized by a certain sparsity and normalized frequency thresholds. This is a simple method, contrary to state-of-the-art methods, which involve machine learning algorithms, as mentioned in the introduction. We tried to find candidate expressions in the Google Syntactic Ngrams dataset[1], using these thresholds. By Looking at the MWE list (Kulkarni & Finlayson, 2011), we set the minimum normalized frequency threshold to that of the least mentioned MWE. In the same manner, we set the threshold of maximum sparsity to the sparsity of the MWE that was mentioned in the corpus across the smallest number of years. Next, we compared three criteria for selecting candidate expressions from the list of collocations (1) by their top trend statistic and normalized frequency, (2) by their top normalized frequency only, or (3) by their lowest sparsity. For each criterion, we calculated *precision@k* by manually labelling *k* top ranking collocations as expressions or not.

## *3.2. Trend Detection*

### 3.2.1. Histogram preparation

Before we could test whether an increasing or decreasing trend exists, we first prepared the MWE histograms for analysis as follows: For each found MWE in the corpus or dataset, (1) we prepared a frequency histogram that detailed the number of mentions of the expression, for each year. Then, (2) we normalized the frequencies, by dividing each yearly frequency by the number of words in the corpus or dataset, for that year. In addition, (3) we segmented the histograms to periods of 7 years, by summing the

[1] http://commondatastorage.googleapis.com/books/syntactic-ngrams/index.html, Version 20130501.

normalized frequency of 7 years, and (4) smoothed histograms by using a moving average with equal weight and size 5 (2 data points on each side).

### 3.2.2. Trend testing

In order to find an increasing or decreasing trend in the usage of expressions, we looked for a test that had the least assumptions about the expression's time-series data. Since counts in the time-series were segmented and smoothed, the assumption of sample independence, used by analysis of variance (ANOVA), for example, could not be assumed, and require adjustments, as suggested by Helsel & Hirsch (2002). Hence, we looked for a different way. Here is a short passage that describes the advantages and disadvantage of non-parametric tests:

> "The main advantages of the non-parametric procedures over parametric alternatives are that the procedures can be used without making too many assumptions about the underlying concentration distributions and. in many cases, their relative simplicity. One disadvantage of these approaches is the relatively low power (i.e. a low probability of detecting a trend) in cases where the assumptions for a corresponding parametric test are reasonable. Another disadvantage for *some* of these procedures is that the non-parametric test may only be able to determine whether a statistically significant trend exists (trend detection) and cannot determine the size of the trend" (Stoeckenius, Ligocki, Cohen, Rosenbaum, Douglas, 1994, p. 45).

We found two tests that answered our needs, in order to test for a trend existence: Kendall's $\tau$ nonparametric correlation coefficient, and Daniels test for trend. Both tests are based on non-parametric correlations that require less assumptions about the dis-

tribution of the data within the time-series than parametric tests: “Kendall’s τ correlation coefficient is a nonparametric correlation coefficient that is often used when distributional assumptions of the residuals are violated or when there is a nonlinear association between two variables.” (Gray, 2007, p. 29) The null hypothesis of Kendall's τ is that there is no trend ($H_0$: $\tau=0$), and the alternative hypothesis is that there is a trend ($H_1$: $\tau\neq0$).

Since the values in a time-series are ordered by time, let $G_i$ be the number of data points after $y_i$ that are greater than $y_i$. In the same manner, let $L_i$ stand for the number of data points after $y_i$ that are less than $y_i$. Given this, Kendall's τ coefficient calculated as

$$\tau = \frac{2S}{n(n-1)} \quad (1)$$

where *S* is the sum of differences between $G_i$ and $L_i$ along the time-series:

$$S = \sum_{i=1}^{n-1}(Gi - Li) \quad (1.1)$$

The test statistic *z* is calculated by

$$z = \frac{\tau}{\sqrt{2(2n+5)/9n(n-1)}}; \quad (1.2)$$

When *n* is large (e.g., *n*>30), *z* has “approximately normal distribution”, so a p-value can be based on the normal distribution table. For smaller *n* values, other tables can be used to get a p-value (Gray, 2007).

Daniels test for trend (Daniels, 1950, as mentioned in U.S. Environmental Protection Agency, 1974) uses Spearman's ρ rank correlation coefficient, which ranks each data point $X_i$ in the time-series as $R(X_i)$. After ranking, ρ is calculated as

$$\rho = \frac{\sum_{i=1}^{n}[R(X_i)-i]^2}{n(n^2-1)}. \quad (2)$$

As with the Kendall's τ correlation test, Daniels test compares Spearman's ρ to a critical value, set by the sample size *n*: When $n<30$, the critical value $W_p$ for a desired p-value is set according to a dedicated table (see table 2, in U.S. Environmental Protection Agency, 1974). When $n\geq30$, the critical value is calculated using $X_p$, which is the p quantile of a standard normal distribution:

$$W_p = \frac{X_p}{\sqrt{n-1}}. \quad (2.1)$$

For example, for a desired p-value>.005 (the equivalent of setting confidence level α=.01), if $n>30$, $W_p=2.5758$, according to the normal distribution tables. If Spearman's $|\rho|>W_p$ a trend exists (U.S. Environmental Protection Agency).

### *3.3. Trend analysis*

A hypothesis of this work is that some synonym expressions can be found by finding expressions with opposing trends. This hypothesis is based on the assumption that each use of a word or expression fulfills a function; if its functionality level is decreased, then it may be compensated by other, synonymous words or expressions that carry the same function unless there is a decrease in the function itself (e.g., due to social or technological changes). We tried to do so by pairing expressions with medium-high negative Kendall's τ rank correlation between them, and manually examining these pairs in order to find synonyms.

In addition, we ordered the list of found trends by the statistic (Kendall's τ) and reviewed the top 30 expressions with increasing trend, and the 30 expressions with

decreasing trend. Then, we looked these expressions in Oxford Historical Thesaurus[2] and tried to find expression synonyms and their dating. This allowed us to compare trends of synonym expressions.

## *3.4. Limitations*

Synonym expressions were found manually, in a thesaurus, which makes it hard to find relationship between opposing trends of expressions with the same meaning. Our method of finding synonym expressions by finding MWEs with negative correlating trends may have been too naïve approach, and could be improved. We also did not look for single-word synonyms or replacements, for MWEs that we found.

Contrary to automatically extracting expressions from the text by co-occurrence measures – or other MWE extraction methods – we used the ready-made MWE list (Kulkarni & Finlayson, 2011) since we were more interested to find synonym MWEs with opposing trends rather than deal with finding the MWEs, first hand.

In addition, the Google Syntactic Ngrams dataset already contains co-occurrence in a summarized form, which made the task of finding co-occurrences a bit redundant.

[2] Oxford Historical Thesaurus – http://www.oed.com

# 4. Experimental Results

## *4.1. Datasets*

For our purpose, we looked for historical corpora or datasets that contained data over hundreds of years that were downloadable for parsing & research. Finally we found the Corpus of Late Modern English Texts[3] (CLMET3.0), and the Google Books Syntactic-Ngrams dataset[4] (Goldberg & Orwant, 2013).

### 4.1.1. The Corpus of Late Modern English Texts (CLMET 3.0)

The Corpus of Late Modern English Texts, version 3.0 (CLMET3.0) contains 333 books from Project Gutenberg and Oxford Text Archive (over 34 Million words, where book length range from 2,387 words to 1,237,952 words) published between the years 1710–1920. Texts are by 212 British authors who are, or were, native speakers of English. (De Smet, 2005; Diller, De Smet & Tyrkkö, 2011) Originally, the corpus was split into 3 periods of 70 years, but we neglected that splitting, and preferred to split into periods of 7 years, which is the minimum period that leaves no period with 0 words in the corpus.

We used Stanford CoreNLP Toolkit (Manning et al., 2014) for POS tagging and Lemmatization. Tagged sentences were then sent to jMWE library (Kulkarni & Finlayson, 2011) that was used to find MWEs from a predefined list it contained, which is described later.

[3] https://perswww.kuleuven.be/~u0044428/clmet3_0.htm
[4] http://commondatastorage.googleapis.com/books/syntactic-ngrams/index.html, Version 20130501.

In order to normalize the counts of the corpus, we counted words per year in a simple manner, by using space, consecutive spaces, or dash as word separators. Words as “won't” were not split into two words.

### 4.1.2. Google Syntactic-Ngrams Dataset

Since CLMET 3.0 is rather a small corpus (34M words), we also inquired MWE usage using the 1 Million English subset of the Google Books Syntactic-Ngrams dataset (Goldberg & Orwant, 2013). The dataset was constructed from 1 Million English books corpus (Michel et al., 2011), published between 1520 to 2008, which originally contained 101.3 billion words. Unlike CLMET 3.0, each line in the dataset already contains 2–5 n-gram (words) collocations that were found in the 1M English corpus, which were found, in total, at least 10 times. Each collocation details its terms, part-of-speech tagging and syntactic dependency labels, total frequency, and a frequency histogram for years the n-gram was found. For example, here is how a line from the dataset looks like:

employed ***more***/JJR/dep/2 ***than***/IN/prep/3 ***employed***/VBN/ccomp/0 12
1855,1 1856,2 1936,2 1941,1 1982,1 1986,1 1989,2 2001,2

For our research, we only used the “arcs” files of the dataset, which contain trigrams – or “dependency triplets” (Lin, 1998, as mentioned in Goldberg & Orwant, 2013) – two content words and optionally their functional markers. Content words “are meaning bearing elements and functional-markers, […] add polarity, modality or definiteness information to the meaning bearing elements, but do not carry semantic meaning of their own.” (Goldberg & Orwant, 2013). These "partial sentences" were checked against jMWE's predefined MWE list (Kulkarni & Finlayson, 2011), as described above (page 11). A bonus of using this dataset is that it saves the task of parsing the

whole corpus for finding collocations.

Since the jMWE parser relies on part-of-speech tagging to find MWEs, we did not differentiate collocations by their syntactic dependency, and summed histograms with similar part-of-speech (POS) in the dataset, into a single histogram, even though they could have different syntactic dependencies. In addition, due to the special function underscores ("_") have in jMWE, underscore characters or tokens were converted to dashes ("-"); if that was the only character of the token/term, it was ignored.

Total counts of the number of tokens in the corpus were taken from the Google Books 1M English Corpus (Google Ngram Viewer, 2009), and since the dataset is already tagged, we only lemmatized the terms before sending the trigrams to jMWE expression detector, in order to bring the words to their stem form.

## *4.2. Multi-word Expressions Found*

12,181 MWEs were found in the CLMET 3.0 corpus, which were mentioned 1,653,410 times in 839,013 different sentences. The histograms of the expressions were segmented into periods of 7 years, for a total of 30 periods, across the corpus 210 years. Of these expressions, 6,032 had a statistically significant increase or decrease in counts, during the 210 years of the corpus (Kendall's $\tau$ correlation coefficient $|z|>3$; $p<.001$). Most of these expressions had high sparsity (4,473 expressions with sparsity > 34%; e.g. most counts were equal to 0 except 5 periods[5], which span 35 years), where 1,559 expressions had sparsity ≤ 34% (34% was arbitrarily decided as a cut-off point).

[5] 5 consecutive periods with counts greater than 0 may represent a positive frequency/count in a single period that was smoothed over 5 periods using an equal weight centered moving average.

Hence, the expressions were filtered to reflect only the latter expressions, with sparsity ≤ 34%.

Even with relatively low sparsity, positive trends usually happen toward the end of corpus, while negative trends occur when the expression is found in earlier periods within the corpus. Looking deeper into the 1,559 expressions, we observed that only 156 expressions had negative trend (negative Kendall's τ z-score).

For the Google Syntactic Ngrams dataset, we created expression histograms for the years 1701-2008, since only from 1701 there is more than 1 book per year (18 book in 1701, raising to 6,147 books in 2008). As a result, histograms' size was 309 years, instead of 489, before segmentation, and 44 periods, or bins, in the final histograms.

When we examined the dataset, and searched for MWEs in its arcs, or trigram files (see research methods, above, for details), we found 45,759 MWEs (out of 65,450 in the MWE index). 41,366 MWEs had a statistically significant trend – an increase or decrease in counts – over the years (Kendall's τ $|z|>3$ or Daniels Test for trend, where Spearman's $|\rho|>0.392$; $\alpha=.01$), and 4,393 MWEs did not have a trend.

The most frequently used expressions were *of which* and *in case* (5% frequency, or 50,000/Million words, over a total of 30 periods – 210 years), while the least frequently used expressions were *bunker buster* and *qassam brigades* (0.122 times per million words, over a total of 28 years). Figure 2 plots the normalized frequency versus rank of each expression that was found, and shows that Zipf's law (Estoup, 1916, as mentioned in Manning & Schutze, 1999), which states that there is a constant relationship between word frequencies and their rank, fits for most of the expressions that were found:

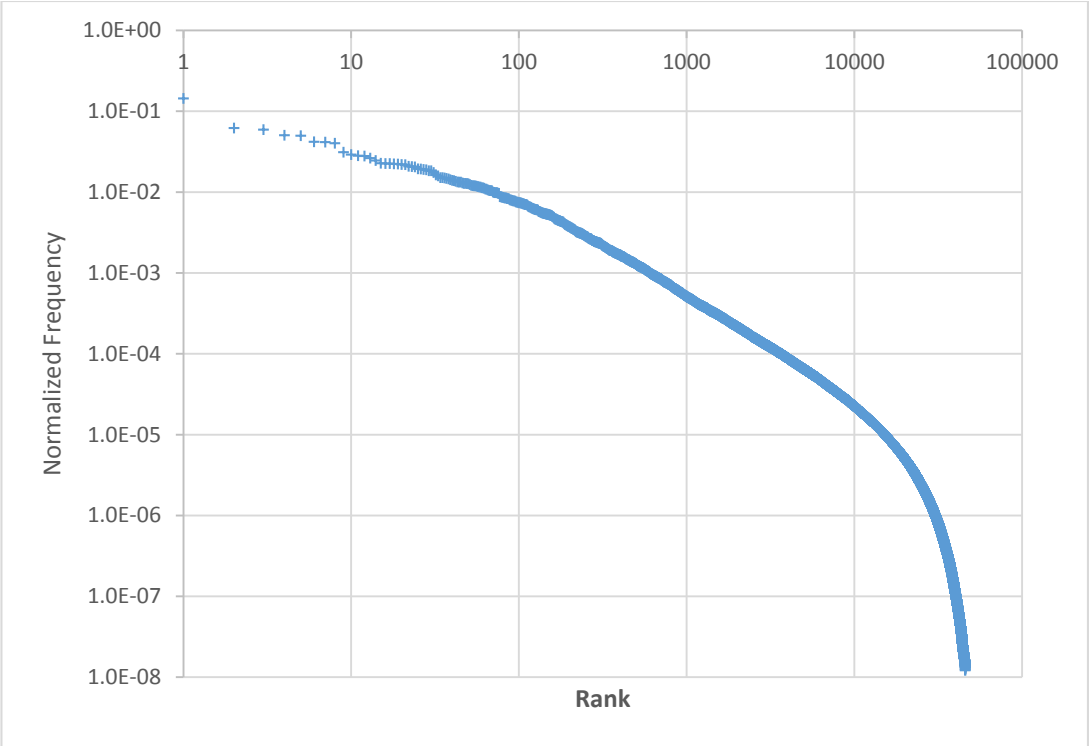


*Figure 2: Rank versus Normalized frequency, using logarithmic scales.*

Many expressions had a sparse histogram, up to 93%; meaning they were used during a rather short period in the dataset (i.e. 90% sparsity corresponds to usage during 4 periods – 28 years). We found that these MWEs were mostly entities, as '*Georgia O'Keef*', though some of them were rarely used MWEs (e.g., *Sheath pile)*, or that they are new expressions or multi-word terms used as an MWE more recently, as *web log* and *pip out*. In order to overcome these "problems", we selected only MWEs with a trend that were used for at least 30% of the time-series periods. That step left us with only 15,895 MWEs (907 with negative trends; 14,988 positive trends) that were widely used across most periods of the dataset, so we could clearly see change in their usage (functionality, or frequency), and focus more on prevalent expressions. Table 3 shows the 30 expressions with the most increasing usage trends, and the 30 expressions with the most decreasing usage trends that were found in the dataset.

### 4.2.1. Finding Candidate MWEs

In addition to finding ready-made MWEs in the dataset, collocations in the dataset that were not recognized as expressions from the ready-made MWEs list (Kulkarni & Finlayson, 2011) were considered *candidate expressions* if they passed two thresholds: We set *normalized frequency threshold* to 1.22E-08, which equals the normalized frequency of the least mentioned MWE that was found in the MWE list (Kulkarni & Finlayson, 2011). This threshold represents 0.122 mentions per million words, or 1,359 mentions across the 111 Billion words in the Google Syntactic n-gram dataset (between the years 1701–2008). We also set *sparsity threshold* to 4 periods – the shortest period an MWE spans, which equals to 28 years. In order to find only newer expressions, we looked for candidate expressions that started to appear since 1904.

Using these thresholds, we found 4,153 candidate expressions. 2,881 of them had a statistically significant trend (α=.01), of which, only 13 showed a decreasing trend. Table 1 shows that 24 (80%) of the top 30 candidate expressions with the most increasing usage trends have MWE characteristics, though some of them are actually multi-word terms used only in a certain profession or subject; For example: *acoustic energy* is probably used in sound engineering studies, *ion[ization] energy* in Physics, while *learning environment* is used in education, and *control subject* in science. However, some of the candidate expressions were not found in dictionaries[6], while showing characteristics of a multi-word expression (marked in bold in Table 1) as *Diary entry*, *older adult*, *entry into force*, *emergency entrance*, etc.

[6] Merriam-Webster dictionary – http://www.merriam-webster.com/dictionary/, and Oxford English Dictionary -- http://www.oed.com/

*Table 1: Candidate expressions with highest, and lowest trends, found in Google Syntactic n-grams dataset. Multi-word expressions not found in dictionaries are bold (p – Proper noun; n – Noun phrase; v – verb phrase; j – adjective; r – adverb; o – other; [none] – statistical collocation that is not an MWE).*

| **30 *candidate expressions* with most increasing trends** | ***[All] candidate expressions* with decreasing trends** |
|---|---|
| ion[ization] energy *(n)* | co. 1922 |
| gastric emptying *(n)* | company 1922 |
| institute of studies *(n)* | reparations commission (n) |
| cultural studies *(n)* | co. 1920 |
| journal of nutrition *(p)* | co. 1919 |
| ultrasonic energy *(n)* | Dalton plan *(p)* |
| **entry into force** *(v)* | Lansing[-]Ishii *(p)* |
| **emergency entrance** *(n)* | company 1919 |
| owned enterprise | company 1921 |
| acoustic energy *(n)* | company 1920 |
| London Faber *(p)* | standpoint of behaviorist |
| however[,] encompass | Leo Nikolayevich [Tolstoy] *(p)* |
| major engagement *(n)* | company 1918 |
| Durham press *(p)* | |
| program encourage | |
| **older adult** *(n)* | |
| policy encourage | |
| **warm environment** *(n)* | |
| Nashville press *(p)* | |
| dysfunction be | |
| **end of runway** *(n)* | |
| learning environment *(n)* | |
| Chapman[-]Enskog *(p)* | |
| correlation function *(n)* | |
| [number]hp engine | |
| control subject *(n)* | |
| **diary entry** *(n)* | |
| **television entertainment** *(n)* | |
| encyclopaedia of Islam *(p)* | |
| intra[-]industry (j) | |

Though *end of runway* is a multi-word term, originally – literally meaning an end of a plane's [landing] runway or a fashion model's [modeling] runway – it could also be

used as an idiom, meaning a *dead end*. The expression *warm environment* is used as a multi-word term, to literally describe a hot environment, but it is also used in a metaphoric way, especially in the field of interior design, to describe a welcoming environment, or space, where one feels comfortable in.

Though we looked for collocations that are mentioned since 1904, it is quite surprising that we did not find seven of the above mentioned expressions in the Merriam-Webster dictionary. This may suggest that the two thresholds that we used could be used to find new candidate expressions, though, we did not compare the method to other existing methods for finding MWEs.

Next, we compared this method to two other methods that select lists of candidate expressions: by taking into account only the normalized frequency values, or the sparsity values, without taking into account the trend value. The different lists are shown in Table 2.

*Table 2: Three lists of candidate expressions, as a result of three different ordering methods.*

| **50 *candidate expressions* with most increasing trends and highest normalized frequency** | **50 *candidate expressions* with highest normalized frequency** | **50 *candidate expressions* with lowest sparsity** |
|---|---|---|
| ion energy (n) | zhou enlai (p) | food engineering (n) |
| gastric emptying (n) | period covered | education program (n) |
| institute of studies | time covered | horror film (n) |
| cultural studies (n) | % of student | save energy (v) |
| journal of nutrition (p) | 5 ht (p) | endemic area (n) |
| ultrasonic energy (n) | campus life (n) | participation in program |
| entry into force (v) | multinational enterprise (n) | older sibling (n) |
| emergency entrance (n) | tropical institute (p) | impact energy (n) |
| owned enterprise | chinese communists (p) | command and school |
| acoustic energy (n) | hague mouton (p) | journal of mammalogy (p) |
| london faber (p) | englewood hall (p) | florida entomologist |
| however encompass | autistic child (n) | food system (n) |
| major engagement | freshmen graduate (n) | sudden urge (n) |

| | | |
|---|---|---|
| durham press (p) | transfer welcome | politics york |
| program encourage | enrollment man | student writing (n) |
| older adult (n) | ion energy (n) | international security (n) |
| policy encourage | blessing from bapu (p) | ward environment |
| warm environment (n) | freshman score | nearby area (n) |
| nashville press (p) | lexical item | memorial center (n) |
| dysfunction be | ethnic enclave (n) | enter job [market] |
| end of runway (n) | laser energy | encouragement and comment |
| learning environment (n) | journal of therapy (p) | entity involve |
| chapman enskog (p) | raven press (p) | skill employ |
| correlation function (n) | j. comp (p) | u.s. energy |
| hp engine | school administrators (n) | shared enterprise |
| control subject (n) | york 1938 | enactment of role |
| diary entry (n) | term care (n) | marginal environment (n) |
| television entertainment (n) | as freshmen | marine environment (n) |
| encyclopaedia of islam (p) | top fifth (p) | thermodynamic entropy (n) |
| intra industry (j) | church related | atherosclerotic plaque (n) |
| psychosocial environment (n) | dairy sci (p) | stanford review (p) |
| state employee (n) | latest week (n) | cultural hegemony |
| socio environmental | energy fig (p) | policy development |
| monetary base | building envelope (n) | successful entrepreneurship |
| zionist enterprise | joint surg[ery] (n) | [on the] ideological end (j) |
| informal english (n) | j. biol (p) | end of exam |
| ensure placement | nature land | end in jail (n) |
| small bowel | gastric emptying (n) | environmental and demographic |
| end of depression | sino soviet | musical ensemble (n) |
| federal employee (n) | chairman mao | economic engagement |
| collective enterprise (p) | york 1941 | gonna end (v) |
| currently engage | j. clin (p) | ensure occur |
| end poverty | soviet relation | ensure relationship |
| institutional context | clin[incal] endocrinol[ogy] (n) | environment and security |
| production engineering (n) | hypertensive encephalopathy (n) | public entity (n) |
| war england | d day (n) | they encapsulate |
| lot of guy | source bureau | violent environment (n) |
| downtown angeles | labor participation | ensure get |
| discrete entity | make serving | coronary syndrome (n) |
| small ensemble | coastal_engineering (n) | picnic table (n) |


*Figure 3: Comparison of the three methods to find candidate expressions, using Precision@k measure. In (a), precision was calculated for all candidate expressions. In (b), precision was calculated after leaving-out proper name expressions (marking them as non-valid expressions).*

Figure 3 shows a comparison of the three methods using *precision@k* measure, which allows to track the precision over a range of candidate expressions (collocations) list sizes: Leaving-out proper name expressions (Figure 3.b), it seems that the best method

is to select candidate expressions by sparsity alone. If we included proper name expressions, there's no clear separation between the methods, and the precision of using normalized frequency alone only slightly outperformed the other two methods for selecting candidate expressions from a list of collocations (Figure 3.a).

## 4.3. Trend analysis

Before looking at expressions with trends, we looked how expressions with no statistically significant trend behave. Figure 4 shows eight expressions that we chose as a representative example.

We chose expressions that have the same range of mentions, more or less, and their Kendall's τ test statistic and Spearman's ρ are relatively far from significant values.

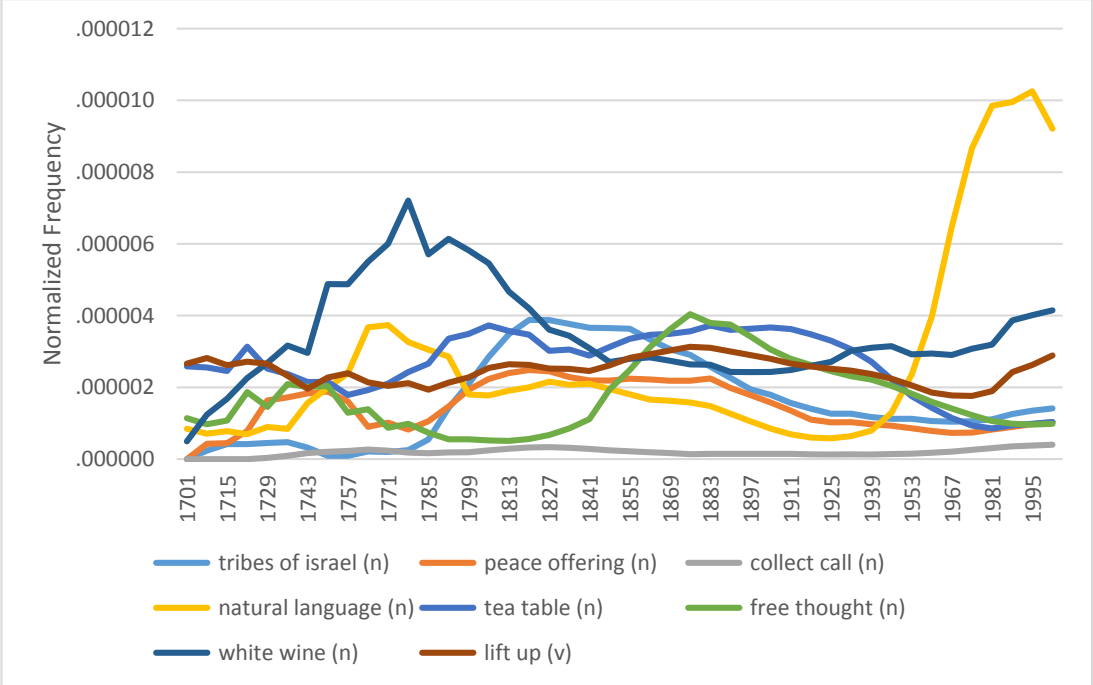


*Figure 4: Sample of MWEs with no statistically significant trend.*

Except *collect call* and *lift up*, which have no trend and seem to behave almost as a straight line, the expressions do not portray a straight horizontal line, as one expects

when no trend is reported. However, this fits our expectations from Kendall's τ and Spearman's ρ to verdict a statistically significant trend only with high confidence (α=.01): Fluctuations in frequency, as the high peak in *white wine*, or *tribes of Israel* are cancelled by previous or future fluctuations, in Kendall's τ formula (1.1), which is based on the sum of differences. *Natural language* may be a trend since 1960, but Kendall's τ coefficient is not "persuaded" so fast that a trend occurs, since the trend is rather short (the last 48 years of over a period 300 years). With these results, we were more confident that our tests are resilient enough, so we did not conduct any further sensitivity tests.

### 4.3.1. Synonyms by negative correlations

After we found expressions trends in the CLMET 3.0 corpus, we tried to find pairs of expressions that have a "mirror image" of their trend. We performed a Kendall's $\tau$ correlation test between the two time-series and for each expression with negative trend, we selected expressions with the most negative correlation, by using -0.6 as a correlation threshold (the closer to -1, the better). Correlating trends for a single MWE ranged from none to 630 other MWEs, and the MWEs with the largest negative correlation were *mean time* vs. *settle down,* with -0.9 correlation.

Examining these negative correlating MWEs for expressions with negative correlating trends, we found only a few synonym expressions, by reviewing the list of negative correlating trends, for each of the 159 expressions that had a negative trend:

- Both *Ill-nature* and *bad temper* refer to a behavioral quality of a person.

- *Common knowledge* and *public knowledge*: *Common knowledge* is "something widely or generally known" ("common knowledge", 2015), and *public* may

also mean "well-known or familiar to people in general" ("public", 2015). *Public knowledge* had a decreasing trend, while *common knowledge* had an increasing trend.

- *Draw off* and *draw back:* Though not perfect synonyms, both phrases can also mean "retreat". *Draw back* had a positive trend within the corpus. Its noun form *drawback* is dated to 1720 ("draw, v.", 2015). The sense of "retreat" is dated around 1300. *Draw off* though, had a negative trend, and in British English also means "to withdraw (troops)" ("draw off [2]", 2015).

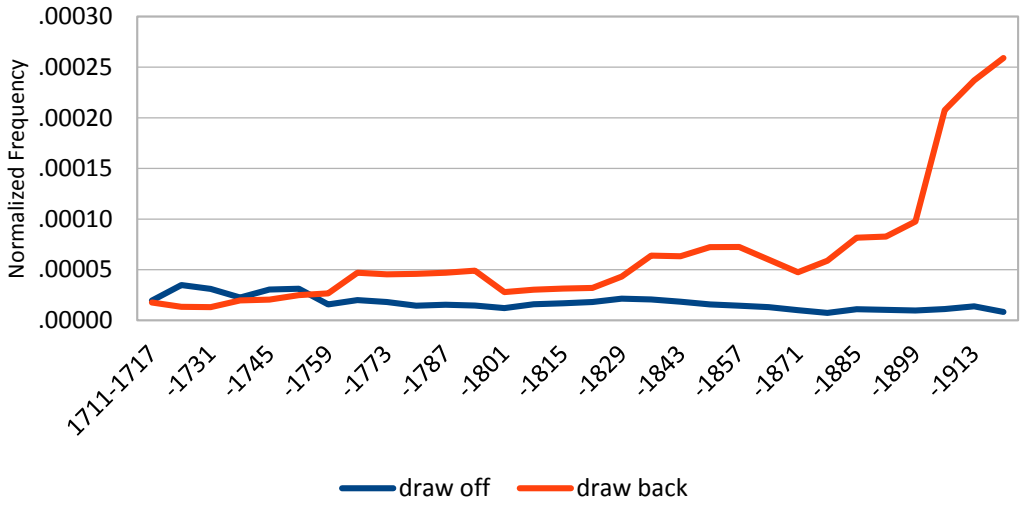


*Figure 5: Comparison between "draw off" and "draw back".*

Following these results, we decided to stop pursuing synonym expressions using this method and continue with analyzing expressions with the most increasing, or decreasing, trends, as described in the following section.

### 4.3.2. Examining Trends

Next, we tried to examine expressions with a trend, and manually find synonyms by using Oxford Historical Thesaurus. We selected the top 30 expressions with decreasing

usage trends, and 30 top expressions with increasing usage trends that were found in the Google Syntactic-Ngrams dataset.

*Table 3: 30 expression with the highest increasing usage trend, and 30 expressions with the most decreasing usage trend. (n – Noun phrase; v – verb phrase; j – adjective; r – adverb; o – other).*

| **Increasing trends** | **Kendall's τ test score** | **Spearman's ρ** | **Decreasing trends** | **Kendall's τ test score** | **Spearman's ρ** |
|---|---|---|---|---|---|
| in turn (r) | 9.568 | 1.000 | take notice *(v)* | -9.184 | -0.994 |
| in practice *(r)* | 9.528 | 1.000 | no more *(r)* | -9.164 | -0.991 |
| better off *(j)* | 9.528 | 1.000 | as much *(o)* | -9.143 | -0.993 |
| think about *(v)* | 9.507 | 1.000 | king James *(n)* | -9.103 | -0.989 |
| work through *(v)* | 9.497 | 0.999 | ill nature *(n)* | -9.062 | -0.990 |
| white woman *(n)* | 9.497 | 1.000 | according as *(j)* | -9.062 | -0.988 |
| human being *(n)* | 9.487 | 0.999 | root out *(v)* | -8.941 | -0.985 |
| talk about *(v)* | 9.487 | 0.999 | think piece *(n)* | -8.799 | -0.987 |
| written record *(n)* | 9.447 | 0.999 | high church *(n)* | -8.718 | -0.979 |
| united kingdom *(n)* | 9.437 | 0.999 | of it *(r)* | -8.718 | -0.976 |
| rule of law *(n)* | 9.406 | 0.999 | make happy *(v)* | -8.658 | -0.979 |
| take into account *(v)* | 9.406 | 0.998 | fourth part *(n)* | -8.658 | -0.965 |
| two dozen *(n)* | 9.396 | 0.998 | St. peter *(n)* | -8.638 | -0.979 |
| rather than *(r)* | 9.386 | 0.998 | church of rome *(n)* | -8.597 | -0.973 |
| go wrong *(v)* | 9.386 | 0.998 | ought to *(v)* | -8.557 | -0.972 |
| human activity *(n)* | 9.376 | 0.998 | good nature *(n)* | -8.557 | -0.971 |
| in fact *(r)* | 9.366 | 0.997 | god almighty *(n)* | -8.536 | -0.975 |
| Cambridge university *(n)* | 9.366 | 0.999 | give ear *(v)* | -8.476 | -0.974 |
| bring together *(v)* | 9.346 | 0.997 | law of nature *(n)* | -8.476 | -0.948 |
| san Antonio *(n)* | 9.335 | 0.998 | let fly *(v)* | -8.415 | -0.973 |
| critical analysis *(n)* | 9.335 | 0.998 | bring forth *(v)* | -8.415 | -0.968 |
| for instance *(r)* | 9.325 | 0.995 | build upon *(v)* | -8.354 | -0.969 |
| end on *(r)* | 9.325 | 0.997 | perpetual motion *(n)* | -8.334 | -0.971 |
| life form *(n)* | 9.325 | 0.997 | revealed religion *(n)* | -8.334 | -0.940 |
| police officer *(n)* | 9.325 | 0.997 | many a *(j)* | -8.314 | -0.968 |
| medical history *(n)* | 9.315 | 0.998 | states general *(n)* | -8.314 | -0.966 |
| run by *(v)* | 9.305 | 0.997 | take care *(v)* | -8.294 | -0.951 |
| conflict of interest *(n)* | 9.305 | 0.998 | as many [as] *(j)* | -8.273 | -0.956 |
| per year *(r)* | 9.295 | 0.996 | take pains *(v)* | -8.273 | -0.940 |
| on and off *(r)* | 9.295 | 0.997 | nemine contradicente *(r)* | -8.253 | -0.957 |

It is noteworthy that some expressions with the most decreasing trends in Table 3 are related to religion (e.g., *revealed religion*, *god almighty*, *Church of Rome*, *St. Peter*, and *high church*). Though our work does not explain language changes, this may be an interesting finding for sociolinguistic researchers, which may indicate a secularization process.

Figure 6 shows the expressions with the 30 most increasing trends, and Figure 7 shows the 30 expressions with the most decreasing trends.

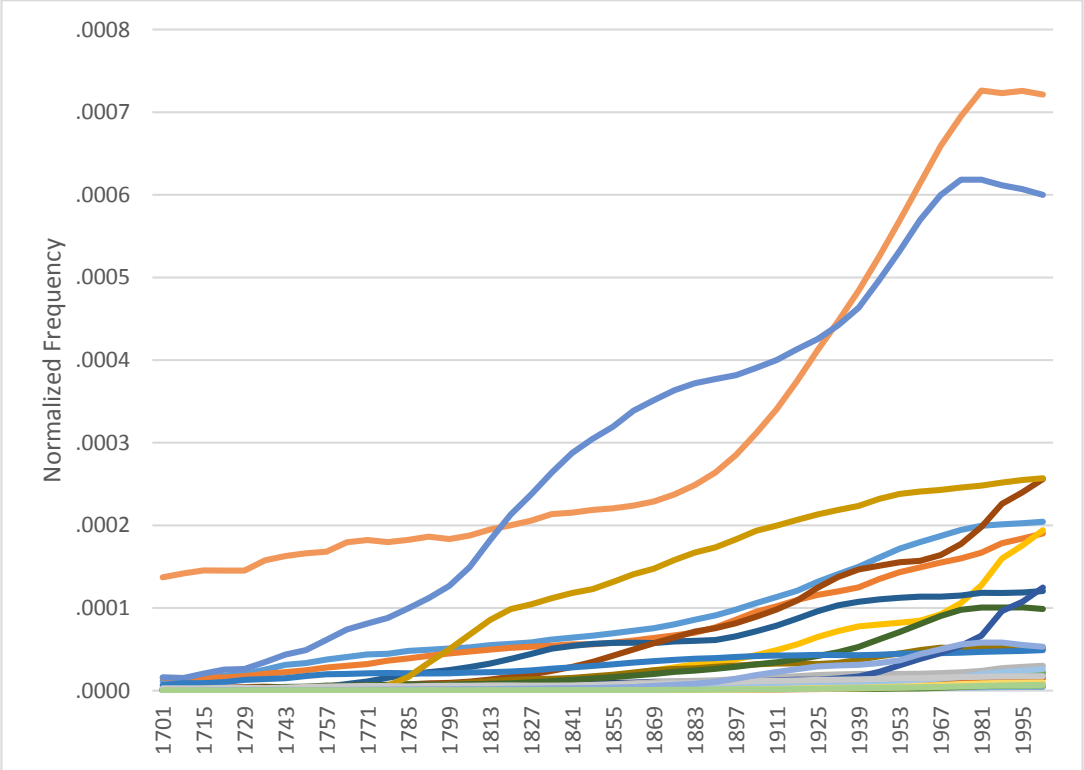


*Figure 6: Most Increasing expression trends, after filtering-out time-series with sparsity >30%.*

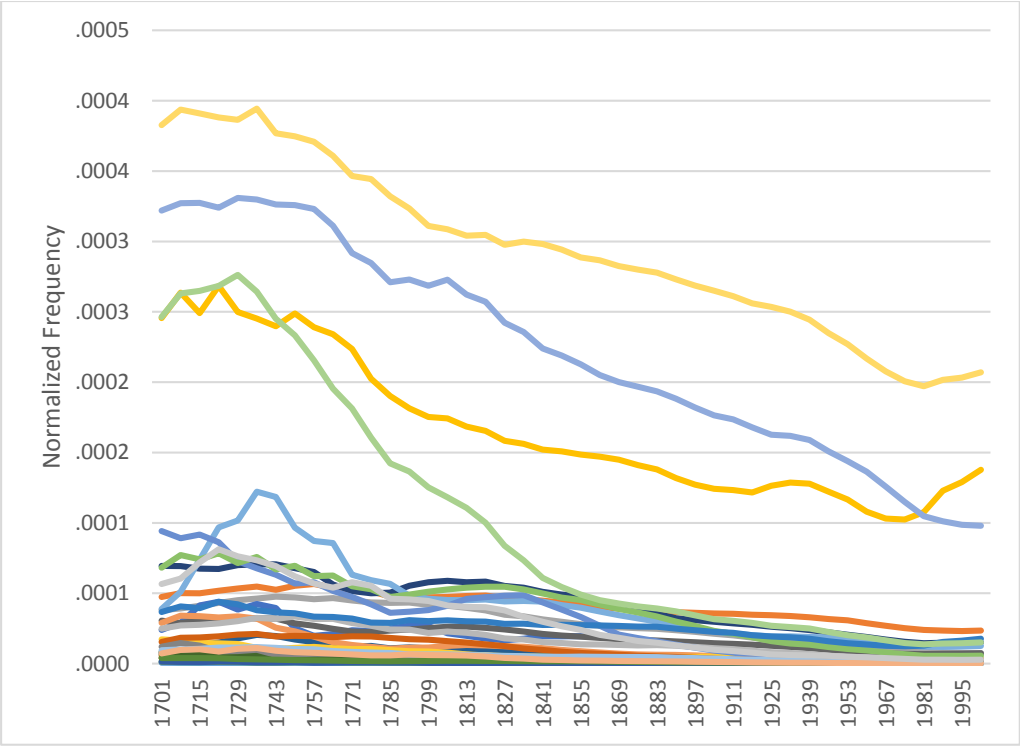


*Figure 7: 30 most decreasing expression trends, without the expression "of it", which has higher levels of frequency, after filtering-out time-series with sparsity >30%.*

### 4.3.3. Top Increasing trends

*In practice*, meaning "In a condition of proficiency in a skill through recent repeated performance […]" is dated to 1816, with a single expression synonym: *one's hand is in* ("in fit condition (for)", 2015). The expression has another meaning, though: "In customary or current use, in vogue; habitually occurring or performed. Obs[olete]", dated to 1537 ("in actual practice", 2015), with synonyms *in deed* (circa 1385), and *in actu* (1902). All synonyms were not found in the ready-made MWE list (Kulkarni & Finlayson, 2011).

*[The] better off* stands for a wealthy person, and is dated to 1895: "With *the* and *pl.* concord. People who are better off financially, considered as a class; = WELL OFF *n.*" ("well-off person or people", 2015).

Earlier synonyms of the expression are *well-to-passer* (1654), *well-to-do* (1829), and *better-to-do* (1860), but none of them were found in the MWE list (Kulkarni & Finlayson, 2011).

*Talk about* is "[…] often used colloq. to contrast something already mentioned with something still more striking; do not talk to me about (something), an exclamation against some new topic of conversation of which one has bitter personal experience" ("talk, v.", 2015). Its synonym expressions are *talk of*, and we compare it also with *speak of* – a synonym not mentioned in Oxford English Dictionary. Figure 8 shows that *speak of* is more widely used than *talk about* since it may have additional meanings, as stating another example to the discussion, where *talk about* and *talk of* are used only to contradict a point in the discussion.

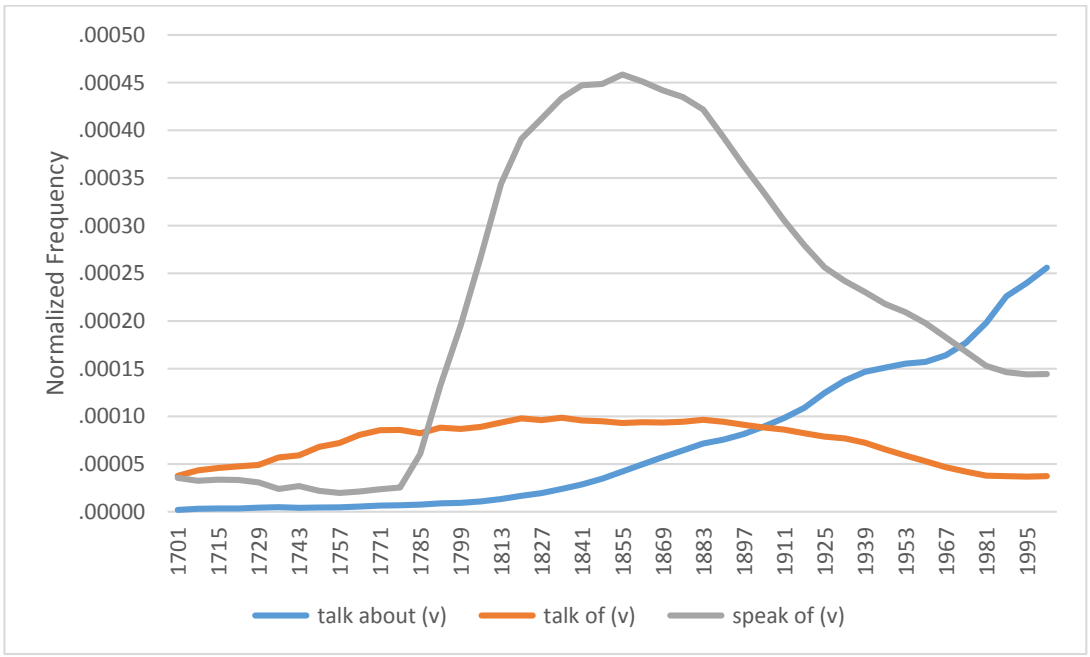


*Figure 8: Comparison of 'talk about' that has an increasing trend of usage, with 'talk of' that has no significant decreasing trend, but shows a decline along the 20th century, and with 'speak of'.*

Figure 9 shows a comparison of the expression *United Kingdom*, compared to *Great Britain*. Though they are not perfect synonyms, we can see a non-significant decline in the usage of G*reat Britain*:

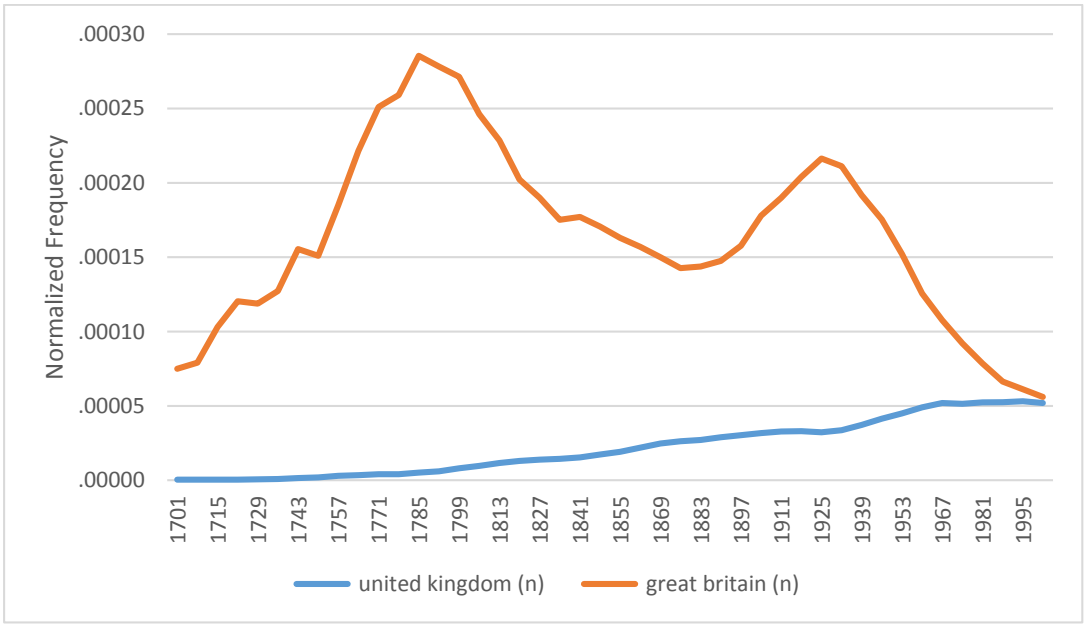


*Figure 9: Comparison of "United Kingdom" that has an increasing trend of usage, with "Great Britain" that has no decreasing trend.*

The expression *[to] go wrong* has several meanings: It could mean to take a wrong way, either literally, in mistake, or morally, by "[…] depart[ing] from moral rectitude or integrity; to take to evil courses" ("wrong, adj. and adv.", 2015). It could also mean that an event "[…] can happen amiss or unfortunately", or that something broke-down, when referring to mechanical or technological things. The expression is also used when one "fail[s] in some undertaking or enterprise, or in the general conduct of life.", or when food "[…] get[s] into bad or unsound condition […]" ("wrong, adj. and adv.", 2015). Its synonym expressions are *to have wough*, *to go will*, *to fare astray* (dated to 1849), *to go/walk/run/step/tread awry*, *go astray*, *fall/run into error*, *fall into mistake*, *make a mistake*, *to come off bluely*, *go down the tube* (1975), *to go haywire* (1929), *to break down* (1837), *go bad* (1799), *to go off* (1695).

Figure 10 compares the usage frequency of *go wrong* with synonyms we found in the MWE list (Kulkarni & Finlayson, 2011):

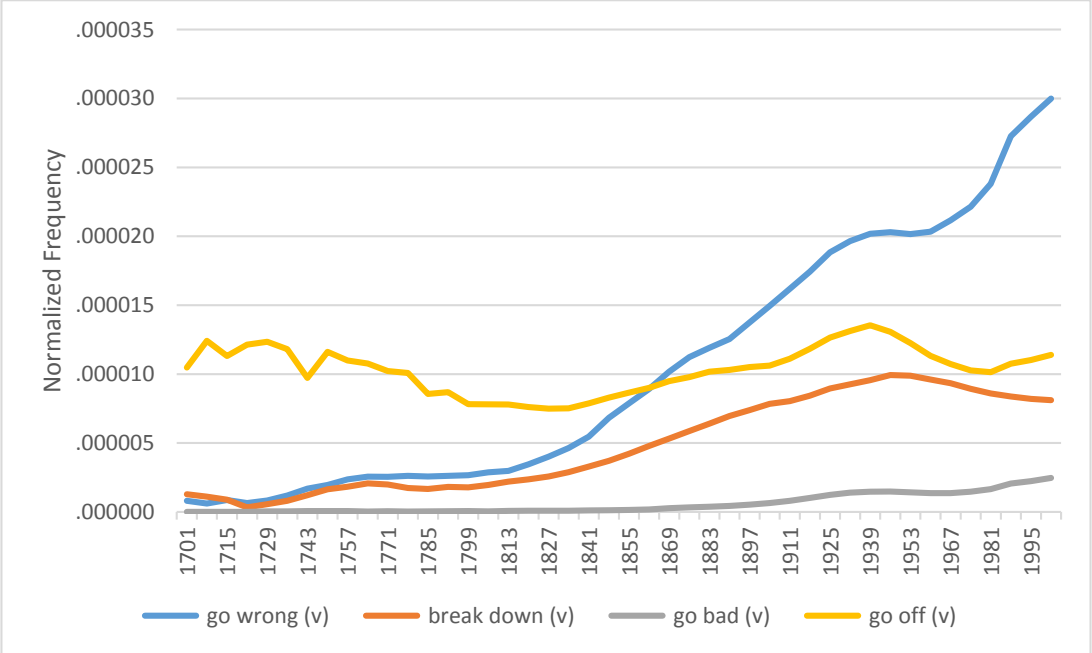


*Figure 10: Comparison of "go wrong" with synonym expressions found.*

*In fact* (dated 1592) is defined as "in reality, actually, as a matter of fact. Now often used parenthetically as an additional explanation or to correct a falsehood or misunderstanding (cf. in point of fact at Phrases 3)" ("fact, n., int., and adv. [P2]", 2015). Its earlier synonym expressions are *in effect*, *in truth* (dated to 1548), *of a truth* (1526), *of truth*, and *for a truth* (1548), *in esse*, *de facto*, *in re[ality]*, *in point of fact*, *in nature*, *in actual fact*, *'smatter of fact* (as a matter of fact; 1922) ("really or actually [adverb]", 2015).

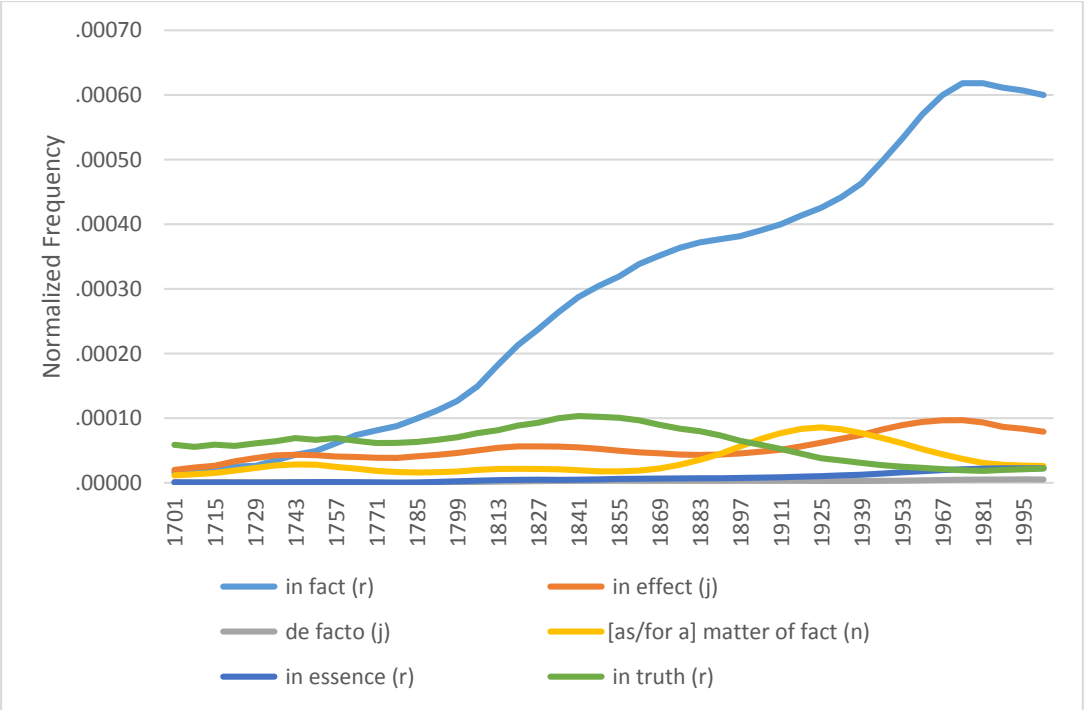


*Figure 11: Comparison of 'in fact' with its near-synonyms.*

*For instance*, as example or an instance of what has been said is dated to 1657. Its earlier synonym expressions are *as namely* (1565), *exampli gratia* (e.g.; 1569), and *for example* (1584). ("for instance or example", 2015). Note that since 1959 *for instance* is also used as a noun: "[...] an example. Freq. in phr. to give (one) a for instance. colloq. (orig. U.S.)." ("exemplifying some rule, activity, quality, etc.", 2015). We found *for example*, but although *for instance* was found in the most increasing trends (Kendall's τ score 9.325), Figure 12 shows the expression *for example* is more trendy, even though its Kendall's τ score is lower (Kendall's τ score 7.525). This could be a result of the Kendall's τ coefficient that considers a monotonic increase a trend, and ignores the amount of increase along the time-series.

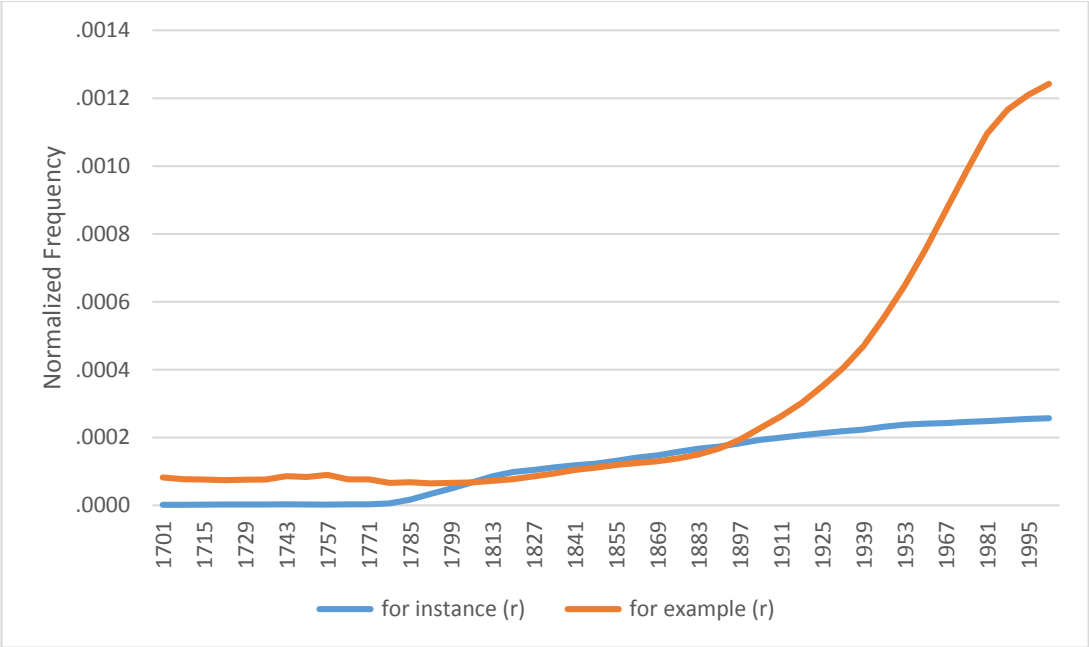


*Figure 12: Comparison of "for instance" and "for example".*

*Police officer* has plenty of synonym expressions: *truncheon officer* (1708), *police constable* (1787), *policeman* (1788), *bow-street officer/runner* (1812), *blueboy* (slang, 1844), *ginger-pop* (1887), *truncheon brearer* (1896), *pavement pounder* (1908), *cow-boy* (1959), *policeperson (1965)*, *smokey bearer* (1974) ("policeman", 2015). We found only *police constable* in the MWE list (Kulkarni & Finlayson, 2011), as shown in Figure 13:

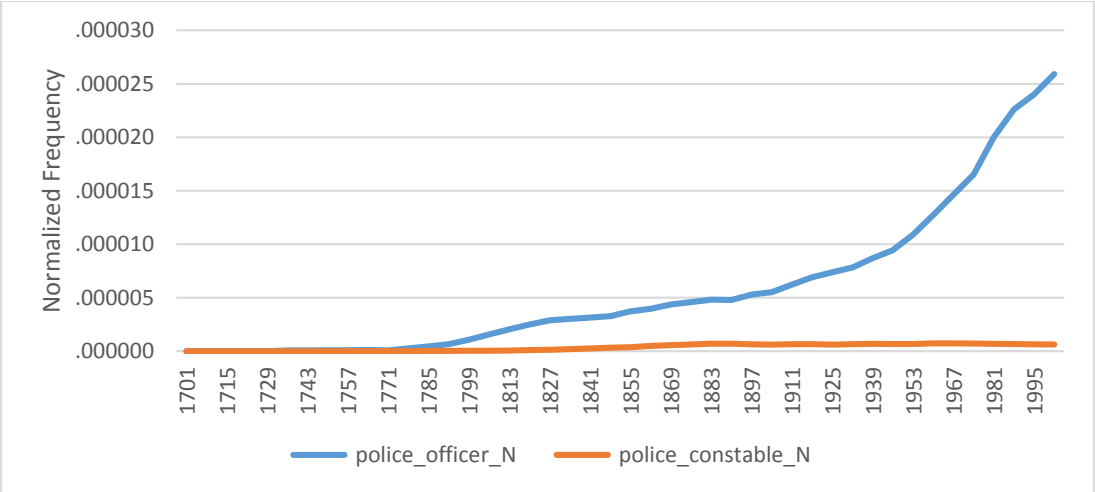


*Figure 13: Comparison of "police officer", which has an increasing trend of usage, with "police constable", which has no evident trend.*

*[To] run by* can either mean to control or manage someone, or a company. In slang, it could also mean to hurt or kill someone by a vehicle ("run, v.", 2015). We tried to look for *manage[d] by* and *control[led] by* as synonym expressions, but did not find them in the MWE list (Kulkarni & Finlayson, 2011).

The expression *on an off* has an earlier synonym expression: *off and on* ("on and off, adv., adj., and n.", 2015), as shown in Figure 14:

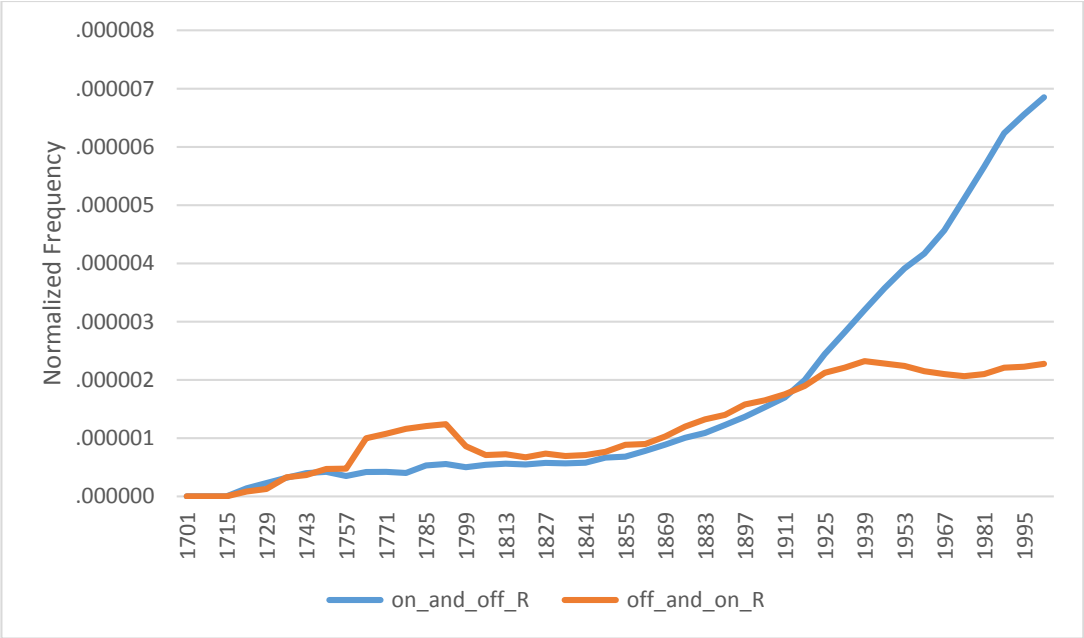


*Figure 14: Comparison between "on and off" and "off and on". Both expressions have statistically significant increase trends, while "on and off" exceeds "off and on" since around 1921.*

We could not find synonym expressions for other expressions, to compare against.

### 4.3.4. Top Decreasing trends

Expressions with decreasing trends were often found in Oxford Online Dictionary[7] as an obsolete, rare, or poetic expressions:

The expression *Many a* is an adjective that describes a large number. It is sometimes "[…] reduplicated for emphasis, as *many and many a*, *many a many*.) […] Now literary" ("many, adj., pron., and n., and adv.", 2015). We did not find its expression synonyms, though, nor its earlier synonym *many one* in the MWE list (Kulkarni & Finlayson, 2011).

[7] http://www.oed.com

*[To] let fly* is a verb that means to attack by shooting missiles, or firing any kind of weapon. Its synonym expressions are *to lay on (circa 1225), to fall on (1387), to set [all] on seven (circa 1400), to give on (1611), to go on (1611), to set on (1670),* and *hop over* ("fly, v.1", 2015). Figure 15 shows other synonym expressions with an increasing trend that may compensate the decreasing usage trend of the expression *let fly.*

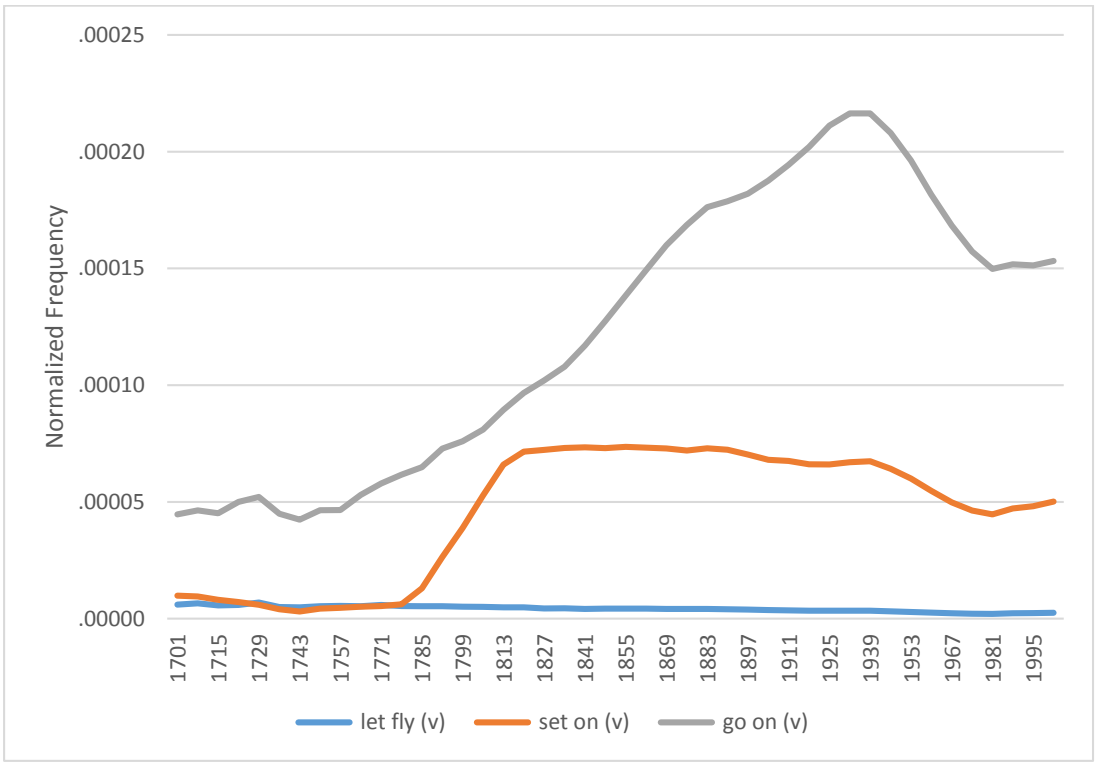


*Figure 15: Comparison of 'let fly' with "set on" and "go on". Set on has an increasing usage trend that is not statistically significant.*

The expressions *take notice* and *give ear* could also be phrased as *pay attention* or *take heed* ("notice, n.", 2015). The expression *pay attention* has an increasing trend, and may partially explain the decrease of *take notice*, as shown in Figure 16. The drastic decrease in usage of the expression *take notice* could also be explained by single-word synonyms, which we did not compare to, as *note*, or *notice*, and *listen*.

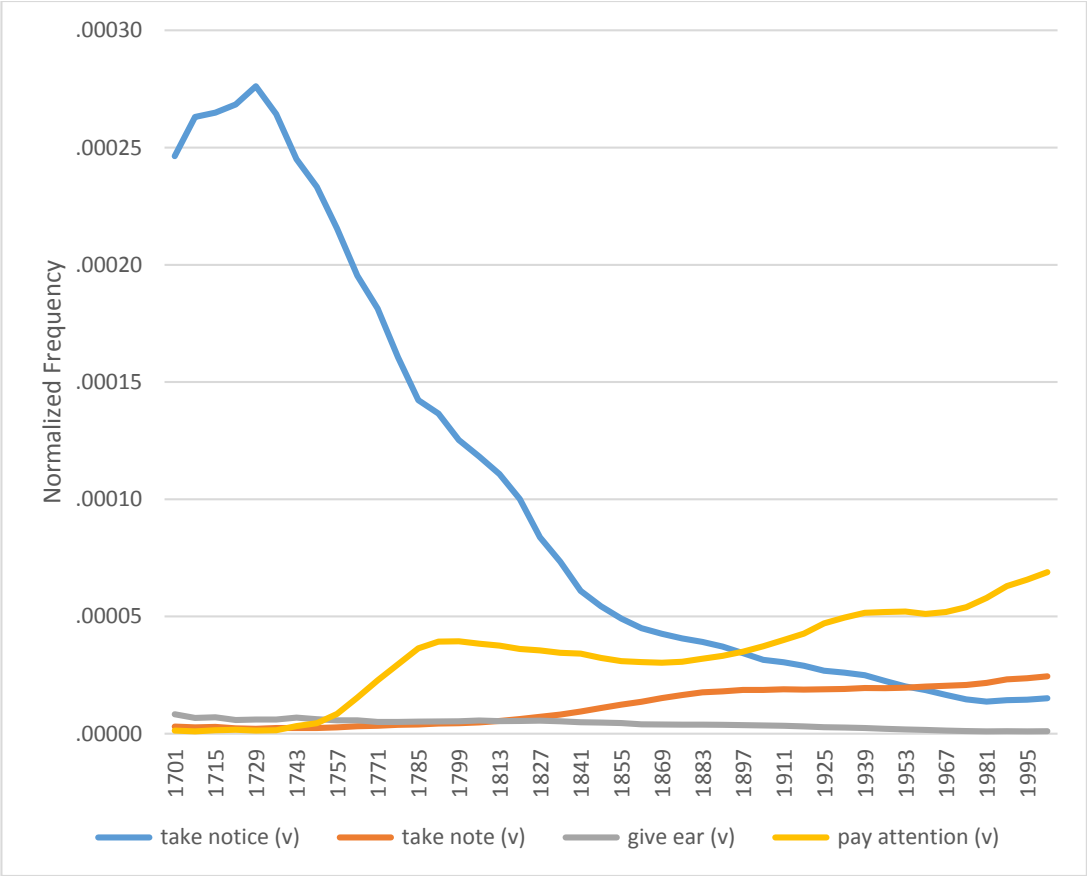


*Figure 16: Comparison of expressions "take notice", "take note", "give ear" and "pay attention".*

The expression *law of nature* is dated to circa 1470, "as implanted by nature in the human mind, or as capable of being demonstrated by reason" ("law, n.", 2015). It is used rarely today because it is used more often today to describe a phenomenon that occurs whenever certain conditions are present, in nature. Its synonyms are *law of kind*, *natural law*, *the law of reason*. The expression *laws of nature* though, in its first meaning, is "[…] viewed as commands imposed by the Deity upon matter, and even writers who do not accept this view often speak of them as 'obeyed' by the phenomena, or as agents by which the phenomena are produced." ("law, n.", 2015). Figure 17 compares the expression *law of nature* with *natural law*:

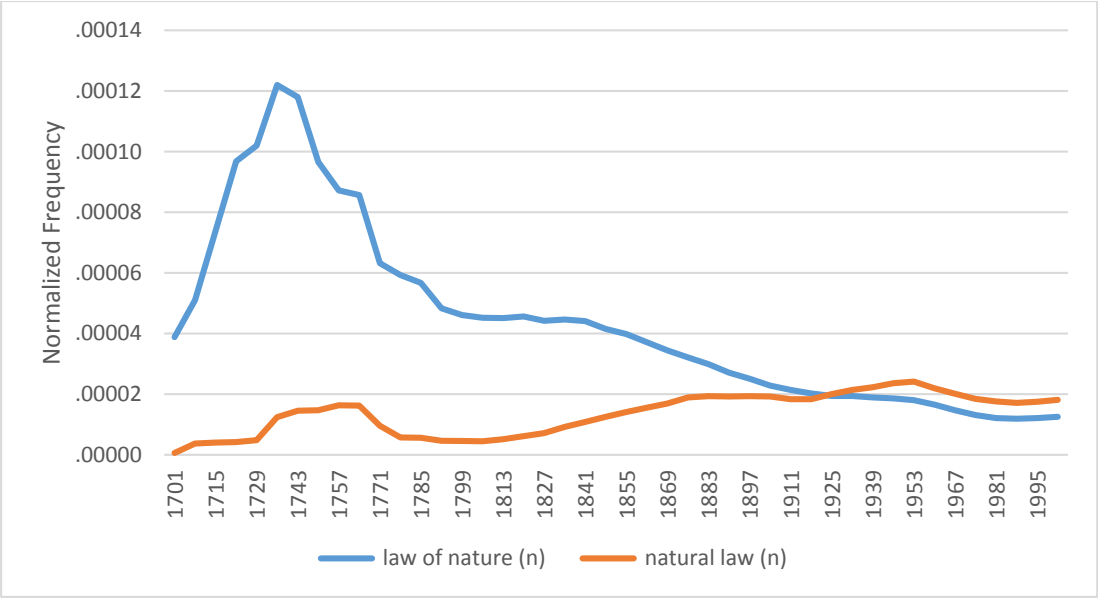


*Figure 17: Comparison of "law of nature" and "natural law".*

*Good nature* as "moral excellence" is dated to circa 1627: "In moral sense: Natural goodness of character; virtue. Obs[olete]. Rare" ("moral excellence", 2015). We did not find any synonym expressions to compare with, but only the single-words synonyms as *virtu* (1906), and *saintliness* (1838). *Good nature* (circa 1450) can mean also "Pleasant or kindly disposition; chiefly denoting a readiness (often excessive) to comply with the wishes or importunities of others, or to permit encroachment on one's rights" ("goodwill or kind intention", 2015). Its synonym expressions are *well-meaningness* (1900), *well-disposedness* (1606), *well-naturedness* (1679), and *well-intentionedness* (1799), but we found none of them in the MWE list (Kulkarni & Finlayson, 2011).

*Ought to* could be use in past tense, which is rare nowadays, i.e., if someone owed or had to pay someone. It could also be used in present tense, where its synonym expressions are *had to*, and *need to* ("ought, v.", 2015). Figure 18 compares *ought to* with the latter synonyms, which show increasing usage trends.

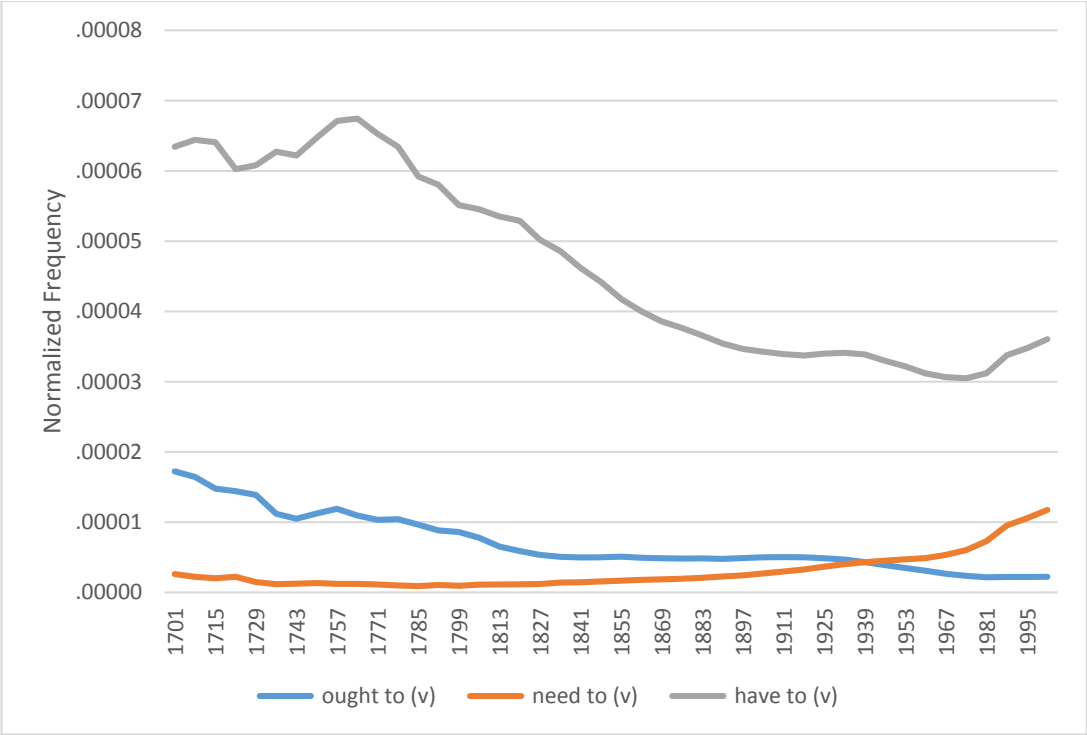


*Figure 18: Comparison of "ought to" with "need to" and "have to".*

*According as* (dated to 1225) is found at Oxford Historical Thesaurus under the sub-category "in conformity with or according to", and it is rarely used: "To the same extent as, in so far as; in proportion as; according as; just as, even as. Now rare except in certain set expressions" ("according as", 2015). Its synonyms are *after* and *thereafter as*, but we did not find them in the MWE list (Kulkarni & Finlayson, 2011).

The expression *no more* has several meanings. In Old English (dated circa 1616) it was used to express that a person is dead, "As predicate: no longer in existence; dead. Now chiefly poet[ically] or ironically" ("dead [adjective]", 2015). Its modern synonyms are *under the daisies*, *dead and gone* and *six feet under*. We found none of the expressions in the MWE list (Kulkarni & Finlayson, 2011). *No more* is also used as adverb, to negate "[…] a second or further alternative: nor yet, neither. Now rare" ("neither", 2015), or used "As a command or request" to stop ("have done with", 2015), with the

synonyms *truce with*, or *truce to*. *No more* as never again, or nevermore, has the synonym *no longer* ("never again", 2015). We found only synonyms for the latter meaning of *no more* in the MWE list (Kulkarni & Finlayson, 2011): *never again* and *no longer*, as shown in Figure 19:

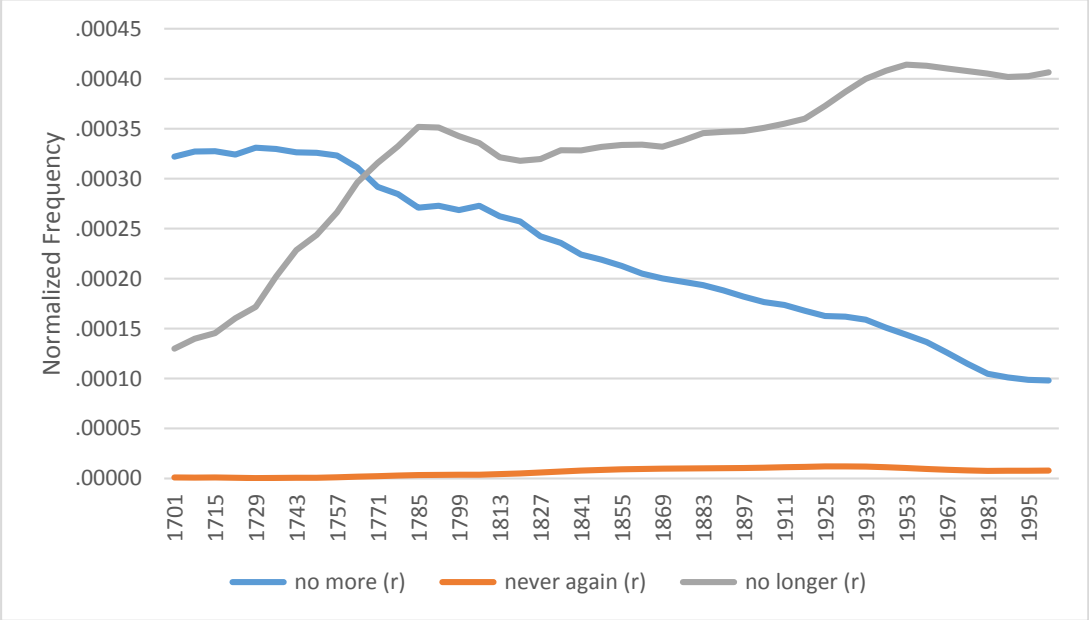


*Figure 19: Comparison of "no more", "no longer", and "never again".*

We could not find synonym expressions for other expressions, to compare against.

## 4.4. Overall expression usage

In addition to finding trends, we summed all normalized histograms, to sense how all expressions behave within the corpus, in average. Contrary to an intuition, we found a statistically significant positive trend in CLMET 3.0 corpus (CLMET: Kendall's τ z-score: 4.993) and Google Syntactic Ngrams dataset (Kendall's τ z: 8.621; $p<.01$), as shown in Figure 9.

The positive trend may be caused by the recency of MWEs in the index, which naturally lacks older MWEs, as it was made during 1–2 decades ago. In other words, it seems there are more recent MWEs in the index, rather than older MWEs.

A comparison of MWE usage during 1711–1920 reveals how Google Syntactic Ngrams dataset is more encompassing than CLMET corpus, resulting in a more stable line (both histograms are segmented into 7 year periods & smoothed).

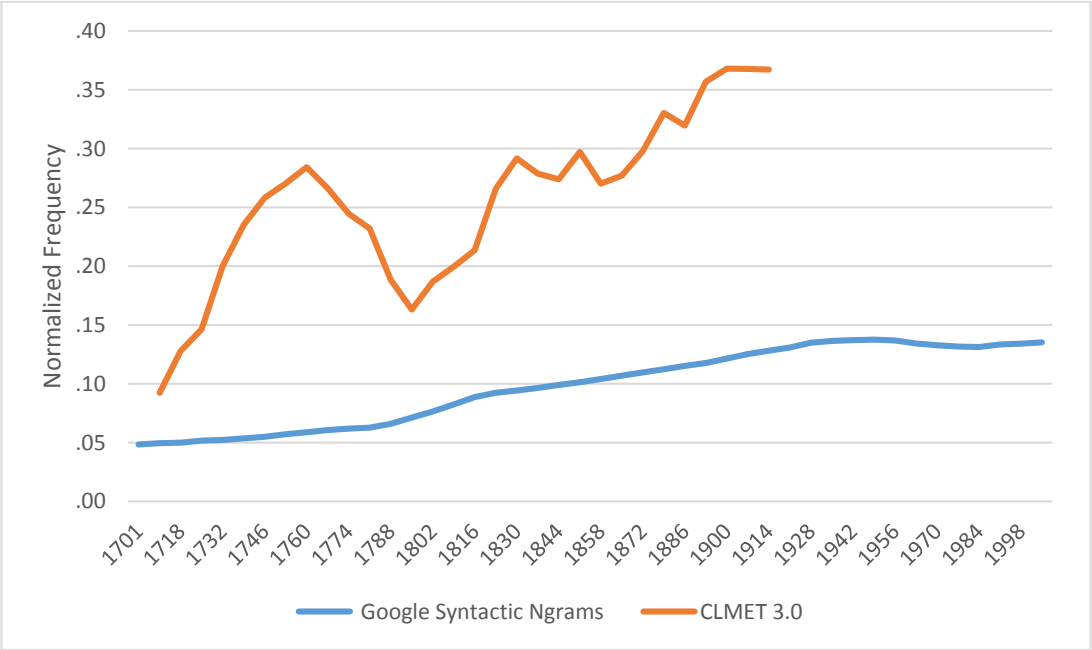


*Figure 20: Comparison between CLMET 3.0 corpus and Google Syntactic Ngrams dataset. Normalized Frequencies are a sum of all MWE normalized frequencies, for a period.*

# 5. Discussion & Conclusions

During this work, we explored the change in Multi-word expressions (MWEs) usage, or functionality over the years. We hypothesized that some synonym expressions can be found by pairing expressions with opposing trends. Examining the pairs resulted in only three near-synonym expressions, out of 159 expressions that had the most decreasing usage trend. These poor results may stem from the fact that the list of predefined MWEs that we used is not complete with respect to language anachronisms, as well as that there may be single-word synonyms that replace multi-word expressions, which we did not examine. Another possible explanation is that the correlation threshold we used for selecting candidate expression pairs (-0.6) was below the actual correlation between the usage trends of some synonymous expressions. Though we found synonym expressions in a thesaurus and not automatically, as we planned, we were still able to "target" on expressions that showed the most remarkable change in their frequency, by looking at their trends (increasing and fading synonymous expressions). Though not in our research scope, we found that some expressions with the most decreasing trends are related to religion, which might interest sociolinguists.

We found that expressions in the MWE list (Kulkarni & Finlayson, 2011) were mentioned in the Google Syntactic-Ngrams dataset for at least 28 years in a row, and showed, as we hypothesized, that it is possible to find *candidate expressions* from collocations using either normalized frequency or sparsity. Using normalized frequency was better, on average, as a method to select any type of candidate expressions, while using sparsity was better, if we are not interested in proper name expressions. A few *candidate expressions* were found to be metaphoric phrases. This may suggest a min-

imum period linguists could test an expression against, before they qualify an expression as a suitable entry in a dictionary, lexicon, or thesaurus. Though MWEs contain various phenomena, and it is hard to create a fully-automatic method that captures them all, we used this threshold along with a minimum frequency threshold to find candidate expressions. We showed that it is possible to find new expressions, not present in a dictionary, even without sophisticated models, though other methods could better solve the issue of separating MWEs from MWTs, and from other collocations.

We demonstrated in a visual way how different corpus or dataset size affects the "image" of the overall MWE usage in language, and that it is important to perform corpus-linguistic research on the largest available corpus. Aggregating all expression counts over each period may suggest that MWEs are on the rise, though this "image" is probably misleading since MWEs are not equally distributed across all periods.

In the future, it is possible to use a method that fine tunes the test for trends, so a short trend towards the end of the time-series would also be recognized as statistically significant. For example, tweaking Kendall's $\tau$ coefficient, especially formula (1.1), to gain weight along the time-series, may boost influence of latter data points.

Future work may also improve the methods for finding MWEs by introducing flexibility in the expression structure, and by using synonym words replacements in order to find variations and synonym expressions. These would assist a lexicographer and language learners to examine usage trends and track language while it continues to evolve.

A collocation trend may also be used as a feature in a learning algorithm that extracts MWEs; the historical perspective of an expression usage has some value for identifying stable expressions, while filtering out short-term keywords and multi-word terms.

# 6. References

According as. (2015). *Oxford English Dictionary*. Retrieved Feb, 2015, from http://www.oed.com/view/th/class/113160

Aitchison, J. (1991). *Language change: Progress or decay?* (3rd ed.). Cambridge, England: Cambridge University Press.

Al-Haj, H., Wintner, S. (2010). Identifying multi-word expressions by leveraging morphological and syntactic idiosyncrasy. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, August 2010, (pp. 10–18), Beijing.

Alippi, C., & Roveri, M. (2006), An Adaptive CUSUM-based Test for Signal Change Detection, *ISCAS 2006*, IEEE.

B. T. Atkins, S., & Rundell, M. (2008). *The oxford guide to practical lexicography*. Oxford: Oxford University Press.

Baker, P. (2011). Times may change, but we will always have money: Diachronic variation in recent British English. *Journal of English Linguistics, 39*(1), pp. 65-88.

Baldwin, T., & Su Nam, K. (2010). Handbook of Natural Language Processing. In N. Indurkhya & F. J. Damerau (Eds.), *Processing* (2nd ed., pp. 267–292). Boca Raton, USA: CRC Press. doi:10.1038/nbt1267

Belica, C. (1996). Analysis of temporal changes in corpora. *International Journal of Corpus Linguistics 1*(1), pp. 61-73.

Buerki, A. (2013). Automatically Identifying Instances of Change in Diachronic Corpus Data. *Presentation at the 'Corpus Linguistics 2013' conference*, 22 to 26 July 2013, Lancaster University, Lancaster (UK)

Calzolari, N. et al. (2002). Towards best Practice for multiword expressions in computational lexicons. *In Proceedings of the third LREC (LREC 2002)*, Las Palmas, Canary Islands, Spain.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Collier, A., Pacey, M., & Renouf, A. (15-16 Aug, 1998). Refining the automatic identification of conceptual relations in large-scale corpora. In *Proceedings of the Sixth Workshop on Very Large Corpora, COLING-ACL*, (pp. 76-84). Montreal.

Common knowledge. (2015). *Collins English Dictionary*. Retrieved March 14, 2015, from http://www.collinsdictionary.com/dictionary/english/common-knowledge

Daniels, H.E. (1950). Rank correlation and population models. J. Roy. Stat. Soc. B12, pp. 171–181.

Dead [adjective]. (2015). *Oxford English Dictionary*. Retrieved Feb, 2015, from http://www.oed.com/view/th/class/16648

De Smet, H. (2005). A corpus of Late Modern English. ICAME-Journal.

Diller, H., De Smet, H., Tyrkkö, J. (2011). A European database of descriptors of English electronic texts. *The European English Messenger 19*(2), 21–35. https://perswww.kuleuven.be/~u0044428/clmet3_0.htm

Draw, v. (2015). *Oxford English Dictionary*. Retrieved March 14, 2015, from http://www.oed.com/view/Entry/57534#eid848052398

Draw off [2]. (2015). *Collins English Dictionary*. Retrieved March 14, 2015, from http://www.collinsdictionary.com/dictionary/english/draw-off

Estoup, J. B. (1916). *Gammes Stenographiques*, 4th edition. Paris.

Exemplifying some rule, activity, quality, etc. (2015). *Oxford English Dictionary*. Retrieved Feb, 2015, from http://www.oed.com/view/th/class/112547

Fact, n., int., and adv. [P2] (2015). *Oxford English Dictionary*. Retrieved March 10, 2015, from http://www.oed.com/view/Entry/67478#eid4939182

Fazly, A., Cook, P., & Stevenson, S. (2009). Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics 35*(1), pp. 61-103.

Fellbaum, C. (Ed.). (1998). WordNet: An Electronic Lexical Database. *Language, Speech and Communication*. MIT Press.

Firth, J. R. (1951/1957). *Papers in Linguistics*. Oxford: Oxford University Press.

Fly, v.1. (2015). *Oxford English Dictionary*. Retrieved March 10, 2015, from http://www.oed.com/view/Entry/72266

For instance or example. (2015). *Oxford English Dictionary*. Retrieved Feb, 2015, from http://www.oed.com/view/th/class/112575

Glynn, D. (2010). Synonymy, lexical fields, and grammatical constructions. Developing usage- based methodology for Cognitive Semantics. In H. J. Schmid & S.

Handl (Eds.), *Cognitive Foundations of Linguistic Usage Patterns*. Berlin: Mouton de Gruyter, 89–118.

Goldberg, Y., Orwant, J. (2013) A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books. In *Proceedings of the second joint conference on lexical and computational semantics (*SEM 2013)*. Retrieved 26 Aug, 2014, from http://commondatastorage.googleapis.com/books/syntactic-ngrams/syntngrams.final.pdf

Goodwill or kind intention. (2015). *Oxford English Dictionary*. Retrieved Feb 2015, from http://www.oed.com/view/th/class/92884

Google Ngram Viewer (2009). Total counts file for English One Million corpus, version 20090715. Retrieved 27 Aug, 2014 from http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-1M-totalcounts-20090715.txt

Gray, K. L. (2007). Comparison of Trend Detection Methods, DOCTORATE. Department of Mathematical Sciences, University of Montana, Missoula, MT.

Have done with. (2015). *Oxford English Dictionary*. Retrieved Feb, 2015, from http://www.oed.com/view/th/class/85913

Heid, U. (2008). Computational phraseology – An overview. In S. Granger & F. Meunier (Eds.), *Phraseology – An interdisciplinary perspective* (pp. 337–359). John Benjamins Publishing Company. Amsterdam/Philadelphia.

Heylen, K., Peirsman, Y., Geeraerts, D., & Speelman, D. (2008). Modelling Word Similarity – an Evaluation of Automatic Synonymy Extraction Algorithms. In Proc.

6th Int. Language Resources and Evaluation, Marrakech, Morocco, 2008, pp. 3243–3249.

Helsel, D.R. and R. M. Hirsch, 2002. Chapter A3. In *Statistical Methods in Water Resources Techniques of Water Resources Investigations*, Book 4. U.S. Geological Survey.

In fit condition (for). (2015). *Oxford English Dictionary*. Retrieved Feb 2015, from http://www.oed.com/view/th/class/85245.

In actual practice. (2015). *Oxford English Dictionary*. Retrieved Feb 2015, from http://www.oed.com/view/th/class/85092

Jackendoff, R. (1997). *The Architecture of the Language Faculty*. Cambridge, USA: MIT Press.

Juska-Bacher, B., & Mahlow, C. (2012). Phraseological change – a book with seven seals? tracing back diachronic development of German proverbs and idioms. In M. Durell, S. Scheible, & R. J. Whitt (Eds.), *volume of Corpus linguistics and Interdisciplinary perspectives on language*. Gunter Narr, Tübingen, Germany.

Kalla, M. (2006). A Diachronic Semiotic Analysis of Words Signifying 'Friendship' in Hebrew. M.A. Thesis, Dept. of Foreign Languages and Literatures, Ben-Gurion Univ. Beer-Sheva, Israel.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory 1*(2), 263–275.

Kulkarni, N., & Finlayson, M. A. (2011). jMWE: A Java Toolkit for Detecting Multi-Word Expressions. In *Proceedings of the 2011 Workshop on Multiword Expressions (ACL 2011)*, pp. 122-124. Portland, OR.

Law, n. (2015). *Oxford English Dictionary*. Retrieved Feb, 2015, from http://www.oed.com/view/Entry/106405#eid39479718

Liebeskin, Dagan & Schler (Aug 8, 2013). Semi-automatic Construction of Cross-period Thesaurus. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 29–35, Sofia, Bulgaria.

Lin, 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual meeting of the Association for Computational Linguistics and 17th international conference on Computational Linguistics – Volume 2, ACL '98*, pp. 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mahlow, C., & Juska-Bacher, B. (2011). Exploring new high German texts for evidence of phrasemes. *Journal for Language Technology and Computational Linguistics, 26*(2), 117–128.

Manning, C. D., & Schütze, H. (1999). Introduction. In *Foundations of Statistical Natural Language Processing* (pp. 23–29). Cambridge, Massachusetts; London, England: The MIT Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguis-*

*tics: System Demonstrations*, pp. 55–60. Retrieved from http://nlp.stanford.edu/software/corenlp.shtml.

Many, adj., pron., and n., and adv. (2015). *Oxford English Dictionary*. Retrieved March 10, 2015, from http://www.oed.com/view/Entry/113819

McCarthy, M., & Carter, R. (2006 [2002]). This that and the other, Multi-word clusters in spoken English as visible patterns of interaction. In McCarthy, M. (Ed.), *Explorations in Corpus Linguistics,* (pp. 7–26). New York, NY, USA: Cambridge University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). Representativeness, balance and sampling. In *Corpus-based Language Studies* (pp. 13–21). London and New-York: Routledge.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: method, theory and practice. Cambridge textbooks in linguistics.* Cambridge. England: Cambridge University Press.

Meusel, R., Niepert, M., Eckert, K., & Stuckenschmidt, H. (2010). Thesaurus Extension using Web Search Engines. In *Proc. Int. Conf. on Asian Digital Libraries*, pp. 198-207, Gold Coast, Australia.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., … Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176–182. doi:10.1126/science.1199644

Moral excellence. (2015). *Oxford English Dictionary*. Retrieved March 10, 2015, from http://www.oed.com/view/th/class/179136

Neither. (2015). *Oxford English Dictionary*. Retrieved Feb, 2015, from http://www.oed.com/view/th/class/82613

Never again. (2015). *Oxford English Dictionary*. Retrieved Feb, 2015, from http://www.oed.com/view/th/class/96637

Notice, n. (2015). *Oxford English Dictionary*. Retrieved March 10, 2015, from http://www.oed.com/view/Entry/128591#eid933873046

On and off, adv., adj., and n. (2015). *Oxford English Dictionary*. Retrieved March 10, 2015, from http://www.oed.com/view/Entry/131310

Ornan, U. (1995). *The Words Not Taken – A Dictionary of Forgotten Words*. The Magnes press, The Hebrew University, Jerusalem, & Schocken, Jerusalem and Tel Aviv.

Ought, v. (2015). *Oxford English Dictionary*. Retrieved March 10, 2015, from http://www.oed.com/view/Entry/133336

Page, E. S. (1954). Continuous inspection scheme. *Biometrika 41*. 100–115.

Philip, G. (2008). Reassessing the canon: 'Fixed' phrases in general reference corpora. In S. Granger & F. Meunier (Eds.), *Phraseology – An interdisciplinary perspective* (pp. 95–108). John Benjamins Publishing Company. Amsterdam/Philadelphia.

Policeman. (2015). *Oxford English Dictionary*. Retrieved March 10, 2015, from http://www.oed.com/view/th/class/174488

Public. (2015). *Collins English Dictionary*. Retrieved March 14, 2015, from http://www.collinsdictionary.com/dictionary/english/public

Ramisch, Carlos. (2013). *A Generic open framework for multiword expressions treatment: from acquisition to applications* (PhD thesis). Universidade Federal do Rio Grande do Sul. Porto Alegre, Brazil, & Grenoble, France.

Really or actually [adverb]. (2015). *Oxford English Dictionary*. Retrieved March 10, 2015, from http://www.oed.com/view/th/class/82683

Rothkegel, A. (2007). Computerlinguistische Aspekte der Phraseme I. In: Burger, Dobrovol'skij, H., Kühn, D., Norrick, P., & Neal, R. (eds.): *Phraseologie. Ein internationales Handbuch der zeitgenössischen*. Forschung. Berlin/New York: Walter de Gruyter, 1027-1035.

Run, v. (2015). *Oxford English Dictionary*. Retrieved March 10, 2015, from http://www.oed.com/view/Entry/168875

Sag, I., Baldwin, T., & Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Mexico City, Mexico, pp. 1–15. doi: 10.1007/3-540-45715-1_1

Sinclair, John. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, (1996). The search for units of meaning. *TEXTUS 9*(1), 75–106.

Talk, v. (2015). *Oxford English Dictionary*. Retrieved March 10, 2015, from http://www.oed.com/view/Entry/197246

Stoeckenius, T. E., Ligocki, M. P., Cohen, J. P., Rosenbaum, A. S., Douglas, S. G. (1994). *Recommendations for Analysis of PAMS Data*. Systems Applications International, San Rafael, California (REPORT).

Tsvetkov, Y., & Wintner, S. (27-31 Jul, 2011). Identification of multi-word expressions by combining multiple linguistic info sources. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)* (pp. 836–845). Edinburgh, Scotland, UK.

U.S. environmental Protection Agency, Office of Air Quality Planning and Standards, Monitoring and Data Analysis Division (1974). *Guideline for the Evaluation of Air Quality Trends. Guideline Series* (OAQPS No. 1.2-014). Retrieved from http://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=9100FOJ5.txt

Well-off person or people. (2015). *Oxford English Dictionary*. Retrieved Feb 2015, from http://www.oed.com/view/th/class/145015

Wrong, adj. and adv. (2015). *Oxford English Dictionary*. Retrieved March 10, 2015, from http://www.oed.com/view/Entry/230802

# תקציר

כדוברי שפת-אם, אנו משתמשים באופן טבעי במונחים וביטויים עדכניים בשפה; נדירים המקרים בהם נבחר להשתמש בביטוי ארכאי. הבחנה זו, בין ביטויים ישנים לביטויים עדכניים, אינה משימה פשוטה ללומדי שפה זרה, ויכולה להועיל גם למילונאים, בבואם להוסיף ביטויים חדשים למילון.

בהינתן קורפוס לקסיקוגרפי, מעניין אותנו להבחין בין ביטויים ישנים לבין ביטויים נרדפים להם שמופיעים מאוחר יותר. אנו מתמקדים בצירופי-מילים (או בפשטות: ביטויים) באורך 2–3 מילים בשפה האנגלית, תוך אפיון כמה תכונות שלהם, כמו תדירות מינימלית ודלילות.

בהנחה שניתן לייצג את מידת השימוש בביטוי (פונקציונליות) לאורך זמן באמצעות סדרה-עתית (time-series), אנו מניחים שלחלק מהביטויים קיים יחס של ישן-חדש – רמז לכך שביטויים נרדפים מחליפים זה את זה. בנוסף לכך , אנו בוחנים את ההשערה שניתן למצוא ביטויים נרדפים על-ידי זיהוי ביטויים בעלי מתאם הופכי בין מגמות השימוש בהם. השערה נוספת שאנו בוחנים היא אם ניתן לאפיין ביטויים בצורה פשוטה בעזרת סף דלילות ותדירות.

כדי למצוא מגמות שימוש בביטויים, חיפשנו ביטויים שהופיעו ברשימת נתונה של 65,450 ביטויים (מבוססת על WordNet 3.0) ב-Corpus of Late Modern English Texts (CLMET 3.0), המכיל 333 ספרים באנגלית שנכתבו בין 1710 ל-1920, וסט הנתונים Google Syntactic Ngrams, המבוסס על מיליון ספרים שנכתבו בין 1502 ל-2008. זיהינו מגמות שימוש בביטויים בעזרת המקדם הלא-פרמטרי Kendall's τ ובעזרת מבחן Daniels, המתבסס על המקדם הלא-פרמטרי Spearman's ρ.

ניסינו למצוא ביטויים נרדפים בעזרת זיווג ביטויים בעלי מתאם הופכי בינוני-גבוה של מגמות השימוש בהם, אך שיטה זו נמצאה כלא יעילה ועברנו לחפש באוצר היסטורי ביטויים נרדפים ל-30 הביטויים בעלי מגמות השימוש החיוביות ביותר ול-30 הביטויים בעלי מגמות השימוש השליליות ביותר. בהמשך, הראינו שקיים לחלק מהביטויים שמצאנו יחס של ישן–חדש.

חישבנו מהנתונים סף תדירות וסף דלילות לביטויים, והשתמשנו בהם כדי למצוא ביטויים מועמדים שמופיעים בטקסטים החל משנת 1904. ניתחנו את התוצאות באופן איכותני.

השיטה למציאת ביטויים נרדפים באמצעות זיווג ביטויים בעלי מגמת שימוש הופכית נמצאה כלא מתאימה: בבדיקה ידנית של זוגות הביטויים, רק מעט מהם היו ביטויים נרדפים. בעזרת אוצר מילים היסטורי מצאנו מספר ביטויים נרדפים לחלק מהביטויים שמצאנו והצגנו השוואה של סדרות-עתיות כדי להראות, כפי שהנחנו, שקיימות מגמות סותרות, או אינטראקציה, בין ביטויים נרדפים.

מצאנו שרשימת הביטויים המוכנה שהשתמשנו בה הכילה ביטויים שהוזכרו לפחות 0.122 פעמים למיליון מילים (1,360 אזכורים מתוך 1.11432E+11 מילים, במאגר הנתונים Google Syntactic-Ngrams) ולאורך תקופה רצופה של לפחות 28 שנים. ערכי הסף האלה יכולים לשמש מילונאים, בטרם יוסיפו ערך למילון או לקסיקון.

## מילות מפתח

צירופי-מילים, ביטויים, ביטויים נרדפים, זיהוי מגמות, שינוי שפה, בלשנות חישובית, בלשנות היסטורית.

**אוניברסיטת בן-גוריון בנגב**
**הפקולטה למדעי ההנדסה**
**המחלקה להנדסת מערכות מידע**

# מגמות שימוש בצירופי-מילים

חיבור זה מהווה חלק מהדרישות לקבלת תואר מגיסטר בהנדסה

מאת: טל דניאל

י"ט תמוז, תשע"ה 6 יולי 2015

**אוניברסיטת בן-גוריון בנגב**
**הפקולטה למדעי ההנדסה**
**המחלקה להנדסת מערכות מידע**

# מגמות שימוש בצירופי-מילים

חיבור זה מהווה חלק מהדרישות לקבלת תואר מגיסטר בהנדסה

מאת: טל דניאל

מנחה: פרופ' מרק לסט

חתימת המחבר..... תאריך.....

אישור המנחה..... תאריך.....

אישור יו"ר ועדת תואר שני מחלקתית..... תאריך.....