This is a preprint version of an article to be published in *Corpora* 16.2 (2021)

An algorithm to identify periods of establishment and obsolescence of linguistic items in a diachronic corpus

Evandro L.T.P. Cunha[1,2,3,] and Søren Wichmann[1,4,5]

[1] Leiden University Centre for Linguistics (LUCL), Leiden University, Leiden, the Netherlands
[2] Department of Computer Science, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil
[3] Faculty of Letters, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil
[4] Laboratory for Quantitative Linguistics, Kazan Federal University, Kazan, Russia
[5] Beijing Language University, Beijing, China
evandrocunha@dcc.ufmg.br / wichmannsoeren@gmail.com

Abstract. When exploring diachronic corpora, it is often beneficial for linguists to pinpoint not only the first or the last attestation dates of certain linguistic items, but also the moments in which they become more strongly established in the corpus or, conversely, the moments in which they, despite still being part of the language, become obsolete. In this paper, we propose an algorithm to assist the identification of such periods based on the frequency of items in a corpus. Our simple and generalizable algorithm can be used for the investigation of any linguistic item in any corpus which is divided into time frames. We also demonstrate the applicability of our method using lexical data from the Corpus of Historical American English (COHA), providing case studies on the statistics and characteristics of words that appear in or disappear from this corpus in different periods.

1 - Introduction

Diachronic and historical corpora are useful tools to study linguistic phenomena that unfold over time, including processes of variation and change. Previous work has employed these kinds of corpora to analyze language change in progress (Hundt and Mair, 1999), to infer cases of variation and change (Bauer, 2002), and to investigate language change using word vector embeddings (Hamilton et al., 2016), to mention a few.

When using diachronic corpora for investigating language variation and change, one of the relevant tasks for researchers is the identification of specific time periods in which certain linguistic items arise and, conversely, vanish. It is particularly valuable to detect when items (i) are first attested, (ii) become established in the corpus, (iii) become obsolete and (iv) are attested last. Although the detection of the earliest and the latest attestation dates of items in a diachronic corpus is trivial, the same cannot be said about their establishment and obsolescence, because there are no clear and commonly accepted criteria for pinpointing when an item is getting established and when it can be regarded as obsolescent (cf. Tichý, 2018).

The aim of this paper is threefold: first, to formulate a set of criteria to define binary notions of establishment and obsolescence of items in a diachronic corpus; second, to present an algorithm to assist the identification of specific time periods of establishment and obsolescence of linguistic items in diachronic corpora according to the previously mentioned criteria; and, finally, to use this algorithm to make a series of general considerations based on real data for the purpose of demonstrating the utility of the methods presented here and for

making some observations on two centuries of the dynamics of the American English lexicon that are interesting in their own right. We will observe, among other findings, that the proportion of words established in a given decade is similar across decades and, by studying the words stemming from different decades that are most frequent today, we will get an impression of how the lexical heritage of contemporary American English bears the imprints of salient aspects of life as it was experienced during specific, previous decades.

The algorithm proposed here is simple and generalizable. It can be applied to any corpus that is divided into time frames, regardless of language or historical period, since it only takes as input information on the frequency of the analyzed items in each time frame. Likewise, the nature of the items under analysis is, in principle, irrelevant to the applicability of our algorithm, so it can also be implemented to examine aspects of language not considered in our case studies, such as phonology or morphosyntax. Moreover, the algorithm, or some derived version, should be generally applicable to the investigation of time series of sociological, anthropological or historical data.

2 - Related work

Previous quantitative investigations on language dynamics have dealt with the notions of birth and death of linguistic items, which are related to the concepts of establishment and obsolescence contemplated here. Petersen et al. (2012), for instance, analyze more than 200 years of data from three different languages with the goal of shedding light on the aggregate dynamics of word evolution in written texts. They investigate variations in the use of words during their lifespans and, among other results, identify a tendency for a peak in word use growth rate to occur around 30-50 years after a word's first attestation in their corpus. Furthermore, the authors find evidence that the dynamics of word evolution might be influenced by historical events, such as wars. This last observation is also made by Bochkarev et al. (2014), who additionally find a relationship between the frequency of a word and its stability in the lexicon of a language, confirming previous results from Pagel et al. (2007). Moreover, Perc (2012) analyzes the evolution of high-frequent English words and phrases, discovering that their lifespan is not uniform across the centuries, and Michel et al. (2011) investigate some patterns in the evolution of English lexicon and grammar.

Certainly connected to the concepts of first attestation, establishment, obsolescence and last attestation of an item in a corpus are the studies that use diachronic corpora to investigate language variation and change. Biber and Gray (2011), for instance, analyze the influence of written language on grammatical change, and suggest that new grammatical uses and functions emerge not only in spoken interaction, but also in written registers. Topics such as the variation of the English genitive (Hinrichs and Szmrecsanyi, 2007), the variation of complex prepositions in Brazilian Portuguese (Shepherd, 2014) and the change in the grammar of English verbs (Hilpert and Mair, 2015), to illustrate, have been considered in previous investigations that made use of diachronic corpora. The use of corpora with the aim of investigating creativity in literary and ordinary language, including novel word formation, is scrutinized by Vo and Carter (2010), while Moon (2010), in tackling the question of what corpora can reveal about lexicon, mentions that these tools might contribute to the analysis of the establishment and the institutionalization of new derivations and compounds in a language.

The notion of establishment of linguistic items in a diachronic written corpus from a particular language is not to be confused with the concept of *entrenchment* of structures in the memory of speakers (Langacker, 1987), which is central in the field of cognitive linguistics. Nevertheless, Schmid (2007) considers that this notion of entrenchment 'also applies to language as such and whole speech communities, because the frequency of occurrence of concepts or constructions in a speech community has an effect on the frequency with which its members are exposed to them' (p. 119). As a consequence, it should be possible to talk of a *degree of entrenchment* of a linguistic item not only in the memory of individual speakers, but also in a specific language. Indeed, Croft (2000) uses the notion of entrenchment in his proposal of an evolutionary model of language change, advocating for a strong relationship between the perpetuation of a given linguistic structure in the language and the degrees of entrenchment of this particular structure in the grammar of speakers. However, in discussing the relationship between frequency in natural language use and the entrenchment of complex linguistic strings in the minds of language users, Blumenthal-Dramé (2012) argues for a weak version of the so-called *corpus-to-cognition principle* — since, according to her, only 'certain corpus-extracted variables may, to some extent, be used as a yardstick for entrenchment in the brain of an average language user' (p. 205). In this study, it is not our goal to contemplate entrenchment in the memory of speakers, nor to elaborate on the relationship between the frequency of linguistic items in a corpus and their entrenchment in the minds of individuals. For this reason, we opted for the use of the term *establishment* and, by not using the loaded term *entrenchment*, we hope to avoid any kind of misinterpretation of the goals of our proposed method.

Regarding the opposite phenomenon — that is, the loss of linguistic items —, Tichý (2018) presents one of the few studies on lexical obsolescence and mortality in English. Using a fine-grained methodology based on the difference between frequency levels in distinct periods of time, the author proposes a method for extracting from large corpora forms that were once common but later became obsolete. Our methodology differs from his in that Tichý is mostly interested in words that were once very common in the language, while our proposed methodology is more flexible in this regard. Also, Tichý's proposal, being more fine-grained, is more computationally demanding, whereas our methodology is simpler and more straightforward. We consider the approaches complementary and imagine that they may even be used together in some specific situation.

Finally, the work of Hilpert and Gries (2009) provides several resources for the assessment of frequency changes in multistage diachronic corpora. The authors present suggestions for the analysis of this kind of data, displaying examples and use cases of great value for historical linguists. In particular, we mention the introduction of the *iterative sequential interval estimation* (ISIE), a method that provides a range of expected frequencies for an item in each time period of the corpus. When the frequency 'happens to go beyond the expected values, we have detected a change that merits further attention' (Hilpert and Gries, 2009: 393).

3 - Defining establishment and obsolescence as binary notions for diachronic corpus linguistics

Dictionaries and glossaries of neologisms (e.g. Ayto, 1989, 1990, 1999; Tulloch, 1991; Algeo and Algeo, 1993; Knowles and Elliott, 1997, to mention works on English) attempt to record recent additions to the language, but their editors are usually aware that what they

characterize as 'new words' might not be new at all. In fact, Tulloch (1991) mentions the potential gap between the point in time in which a word enters the language and the moment when the general public becomes aware of it — which is the occasion when the neologism might be included in the most prestigious dictionaries and can be considered 'established in the language' (Ayto, 1999: iii). Still, most of the past studies mentioned in Section 2 that analyze time periods in which linguistic items arose and vanished associate birth and death with, respectively, first and last attestations in a corpus. In this paper, we argue and show evidence that the first appearance of an item in a corpus may occur considerably earlier than its establishment in the corpus itself and, conversely, that an item might still appear in the data long after it became obsolete (see Section 5). This fact suggests that it may often be convenient to discriminate between first attestation and establishment as well as between last attestation and obsolescence, so as to obtain a more accurate description of the lifespan of a linguistic item.

As pointed out by Widdowson (2000), it is important to emphasize that a corpus is different from a language and, consequently, that the establishment or the obsolescence of an item in a corpus does not necessarily imply its establishment/obsolescence in a language. At most, it might be claimed that a corpus represents part of a language and that a relationship between these two entities exists.

We are interested in defining binary (rather than continuous[1]) notions of establishment/obsolescence in order to indicate whether a linguistic item may have arisen in or vanished from a diachronic corpus during the period covered by it. This is particularly useful for researchers interested in extracting lists of candidate items for further research (see Section 5.3). We stipulate that, in a particular corpus which includes diachronic information, each linguistic item (be it a word, a morpheme, a syntactic structure or other) may usefully be classified as being in one — and only one — of the following possible states in a given period: (a) established; (b) obsolete; (c) permanent; (d) short-lived; (e) random. These states refer to diachronic patterns of appearance of the item through the corpus. The state *established* concerns items that, although not frequent (above a given threshold) in the beginning of the period, rise in frequency at some point and remain frequent until the end of the period covered the corpus. In other words, established items were not part of the language represented by the corpus, but at some point during its time span they flourished and remained frequent afterwards. The state *obsolete*, conversely, refers to items that are frequent (above a given threshold) in the beginning of the period covered by the corpus, but which at some point decrease in frequency. They are, therefore, items no longer in general use by the end of the corpus, although they may linger on as old-fashioned forms or archaisms making occasional appearances. The state *permanent* describes items that are frequent enough through the whole period covered by the corpus. The state *short-lived* regards items that flared up for some time and then, still during the period covered by the corpus, decreased in frequency again. Finally, the state *random* is reserved for items that do not show any of the aforementioned patterns. In the next section, we further develop this categorization by presenting our proposed methodology for classifying items into the above-mentioned classes.

1 In other words, our aim is to provide sets of (candidate) established/obsolete items rather than some sort of 'degree of establishment/obsolescence' per item.

4 - The algorithm

4.1 - Requirements

In order to be accessed by our proposed algorithm, a corpus must be divided into time frames. These time frames might delineate any desired period of time, depending on the nature of the data and on the research goals. Each one of these time frames may represent, for instance, a period of several years, or one decade, or one year, or even one day — the latter in the case of research using data from online social media platforms, for example. For methodological reasons, it is to be preferred that time frames are uniform (both in corpus size and duration, whenever possible) across the whole corpus, but this is not a strict requirement and alternative methods (such as the one proposed by Gries and Hilpert (2008)) could be used to divide the corpus in time stages. Also, our method relies on the use of topically coherent corpora, so as to avoid that changes in sampling across time lead to change in the frequency of linguistic items.

In our method, when the frequency of a given item in a certain time frame is above a definite threshold, it is represented by the digit `1`; when this frequency is below this threshold, by `0`. For example, in a corpus divided into six time frames, the *diachronic sequence* of an item whose frequency exceeds the threshold only in the last time frame is denoted by `000001`, while the sequence of an item whose frequency exceeds the threshold in all but the second and third time frames is denoted by `100111`.

We leave the definition of the boundary between assigning a `0` or a `1` in the diachronic sequence as a choice for the researcher who will use our algorithm, since this depends on additional methodological choices and assumptions. We strongly discourage, however, the use of absolute frequencies as thresholds (as they are dependent on the size of the corpus in each time frame) and, conversely, encourage the use of relative frequencies. For example, a `1` might be attributed to a given item in a particular time frame in case its frequency exceeds *n* % of the total size of the corpus in that time frame; otherwise, a `0` will be attributed. A simple and useful case is when this boundary is set on a really low relative frequency (e.g. 0,00000001% of the corpus size). In this case, the mere presence of the item in the time frame is enough to assign a `1` to it. This simple situation is convenient, practical and might still give interesting results, such as the ones we display on Section 5.

In Section 4.2, we introduce the rules regulating a first algorithm aimed at the categorization of linguistic items into one of the previously mentioned states — *established*, *obsolete*, *permanent*, *short-lived* or *random*. We begin by stating naive rules that are ultimately not satisfactory for our intentions. In Section 4.3, however, an improved version of these rules, more effective for the purposes of the goals declared here, is presented.

4.2 - Rules for a naive algorithm

A first (and naive) version of an algorithm aiming to solve the task of categorizing a linguistic item into one of the aforementioned states may be based on the following rules:

- Established items: those that are not frequent enough in the corpus before a certain time frame, but from a given point start to exceed the frequency threshold in all of the following time frames, without exception. Example of a diachronic sequence in a corpus containing six time frames: `000111`.

- Obsolete items: those that are frequent in the first time frame(s), but from a given point onwards are not frequent enough in any of the following time frames, without exception. Example: `111000`.
- Permanent items: those that are frequent in all time frames, without exception. Only possible diachronic sequence: `111111`.
- Short-lived items[2]: those that are not frequent enough in the extremes of the period covered by the corpus, but that are consistently frequent during an intermediate period. Example: `00111100`.
- Random items: those that do not fit into any of the previous cases. Example: `100101`.

It is clear that these rules only work for what we might call 'perfect' patterns, in which linguistic items 'appear' or 'disappear' at a certain point and keep this status until the end of the period covered by the corpus, without fluctuations. According to this method, an item which, in a corpus divided into ten time frames, exhibits the pattern `0001011111` is considered an example of a random pattern, even though it is obvious for us that it clearly illustrates an item established sometime around the middle of the period covered by the corpus. To solve this issue, an improved version of these rules, allowing for some deviations from perfect patterns, is presented in the next section. Without the allowance of these deviations, the low frequency of an item in a specific time frame would be too severely punished, being enough to disregard the item as an innovation; conversely, the presence of an item in a specific time frame could be enough to disregard it as an obsolete item.

4.3 - Proposed algorithm

Here, we propose an algorithm that enhances the previous approach by allowing for small deviations from perfect patterns, thus making it possible to include more (and more accurate) data into the lists of established and obsolete items of a corpus. The core idea is (i) to compare the observed (real) diachronic sequences of each item in the corpus with perfect patterns for establishment and obsolescence, and then (ii) to select a specific time frame as representing the time of establishment or obsolescence, using the criterion that it should be the time frame that produces the smallest amount of deviation from these perfect patterns.

Consider the following fictitious example. In a corpus divided into ten time frames, the linguistic item *A* exhibits the diachronic sequence `0001110111` — according to which *A* is not frequent enough in the initial periods of the corpus, but after time frame four it is consistently frequent, with the only exception being time frame seven. Our algorithm inspects each position in between two adjacent time frames, starting from position one (which lies in between the first and the second time frames, as in `0_001110111`). The perfect pattern indicating the establishment of an item in this position is `0_111111111` (i.e., the item is not present before the position and is consistently present after it), while the perfect pattern indicating its obsolescence at this point is `1_000000000`. Here, the algorithm investigates the observed sequence for item *A* and counts deviations from the two perfect patterns. By 'deviations' we mean differences in particular points of the diachronic sequences: for instance, if, in a given place, a `0` is found in the observed sequence when a `1` is expected according to the perfect pattern, then we detect a deviation[3].

---

2 We acknowledge an anonymous reviewer for suggesting the inclusion of the category of short-lived items.

3 These deviations might be counted, for example, by employing an edit distance algorithm, such as the

Let us return to the example of the sequence `0001110111`. At the first position, when the assumption is that the item gets established after that point in time, the algorithm finds three deviations from the perfect pattern (the three `0`s in time frames two, three and seven); when the assumption is that the item becomes obsolete, there will be seven deviations from the perfect pattern (the `0` in time frame one and the six `1`s in time frames four, five, six, eight, nine and ten), as illustrated below, where digits in boldface indicate deviations:

Observed sequence:	`0_`**`00`**`1110`**`1`**`111`	Observed sequence:	`0_00`**`1110111`**
Perfect pattern:	`0_`**`11`**`1111`**`1`**`111`	Perfect pattern:	`1_00`**`0000000`**
(establishment)		(obsolescence)	

After these results have been obtained for the first segmentation, the algorithm moves to the next position (`00_01110111`). Here, two deviations from the perfect pattern of establishment (the two `0`s in time frames three and seven) and eight deviations from the perfect pattern of obsolescence (the two `0`s in time frames one and two, and the six `1`s in time frames four, five, six, eight, nine and ten) are found. In the third position (`000_1110111`), only one deviation from the perfect pattern of establishment is found (the `0` in time frame seven), while nine deviations from the perfect pattern of obsolescence are detected (the three `0`s in time frames one, two and three, and the six `1`s in time frames four, five, six, eight, nine and ten). In the next step, two deviations from the perfect pattern of establishment (the `1` in time frame four and the `0` in time frame seven) and eight deviations from the perfect pattern of obsolescence (the three `0`s in time frames one, two and three, and the five `1`s in time frames five, six, eight, nine and ten) are identified in the fourth position (`000_1110111`). The process continues until all positions[4] are analyzed, after which the position producing the smallest number of deviations can be found. This position will represent a possible moment of establishment or obsolescence. In the case of item *A*, Table 1 shows that the smallest number of deviations is found under the assumption of establishment (rather than obsolescence) and is observed in position three, indicating that this linguistic item might have been established in the corpus immediately after this point — that is, within time frame four.

Let us go on to consider, in the same corpus, a linguistic item *B* exhibiting the diachronic sequence `1111100100`. After the inspection of the nine positions in between each two adjacent time frames, the proposed algorithm outputs that the smallest number of deviations from a perfect pattern is found in position five, but now the assumption is that of obsolescence. In this case, the decision implies that the item has become obsolete in the time frame following that position, which corresponds to time frame six, as displayed again in Table 1.

Levenshtein distance algorithm, that returns the minimum number of single-character edits required to change one sequence into the other.

4 Note that the number of positions to be analyzed equals *tf* - 1, where *tf* represents the number of time frames in which the inspected corpus is partitioned.

Table 1. Number of deviations in each position according to the proposed algorithm for two fictitious examples *A* and *B*. In the first example, the smallest number of deviations is observed in position three under the assumption of establishment, indicating that *A* might have gotten established immediately after that position (in time frame four). In the second example, the smallest number of deviations is observed in position five under the assumption of obsolescence, suggesting that *B* may have become obsolete immediately after that position (in time frame six).

Item *A*			**Item *B***		
Position in the diachronic sequence	**Deviations (establishment)**	**Deviations (obsolescence)**	**Position in the diachronic sequence**	**Deviations (establishment)**	**Deviations (obsolescence)**
1 (`0_001110111`)	3	7	1 (`1_111100100`)	5	5
2 (`00_01110111`)	2	8	2 (`11_11100100`)	6	4
3 (`000_1110111`)	**1**	9	3 (`111_1100100`)	7	3
4 (`0001_110111`)	2	8	4 (`1111_100100`)	8	2
5 (`00011_10111`)	3	7	5 (`11111_00100`)	9	**1**
6 (`000111_0111`)	4	6	6 (`111110_0100`)	8	2
7 (`0001110_111`)	3	7	7 (`1111100_100`)	7	3
8 (`00011101_11`)	4	6	8 (`11111001_00`)	8	2
9 (`000111011_1`)	5	5	9 (`111110010_0`)	7	3

It is worth noting that, for each position, the number of deviations from the perfect pattern indicating establishment plus the number of deviations from the perfect pattern indicating obsolescence equals the amount of time frames in the corpus. This is obviously expected, since each `0` or `1` in the observed sequence is always a deviation from a perfect pattern (either regarding establishment or obsolescence), but never a deviation from both.

The proposed algorithm will always output a smallest number of deviations from the perfect patterns, but this value might be considered excessive in some cases. For this reason, a cut-off point of the number of acceptable deviations from establishment and obsolescence should also be defined, and cases that exceed this threshold should be assigned to the pool of cases of random distributions. This cut-off point must be set by the researcher according to some sensible considerations that will vary according to the type of corpus in question: if it is a lexical corpus of child language acquisition with day-to-day recordings, for example, there might be many deviations since a single child is not expected to exercise its full vocabulary every day; if it is a large historical corpus of texts with yearly time frames, the cut-off point could be set to fewer deviations[5]. Here we must refrain from generalization about such thresholds, but we give an example of how to derive one from the behavior of a specific corpus in Section 5. What we mainly want to stress is that the use of an algorithm such as the one described here has the advantage that there has to be such an explicit threshold. Even if it is defined in somewhat *ad hoc* ways in individual cases, it will force researchers to be specific about their choice, enhancing transparency and replicability of a given study.

Finally, in case of ties such that the smallest number of deviations occurs at more than one position, we advocate for choosing the position that includes more time frames with the item being present so as to maximize the amount of positive attestations after minimizing the amount of deviations. An example would be as follows: in a diachronic sequence such as `1111101000`, where the smallest number of deviations from a perfect obsolescence pattern

5 It is trivial to observe that the naive rules presented in Section 4.2 correspond to the algorithm proposed here when this cut-off point equals to zero.

(which is one) is achieved both in positions five and seven, we favor choosing the latter (corresponding to the eighth time frame) as the moment of obsolescence; conversely, in a sequence such as `0001011111`, we favor choosing the fourth time frame (rather than the sixth) as the moment of establishment.

In conclusion to the present section, we present a summary of the steps made by the algorithm.

Summary of the algorithm

1 - Go to the first position in between two adjacent time frames.
2 - Calculate the number of deviations from both the perfect patterns of establishment and obsolescence.
3 - If there are unexplored positions in between two adjacent time frames, go to the next position and repeat step 2; otherwise, go to the next step.
4 - Compare the value found for the smallest number of deviations S with the maximum threshold for deviations allowed T;

I - If $S > T$, the item is considered neither established nor obsolete.
II - If $S \leq T$:

i - resolve potential ties by choosing the position that includes most time frames with the item being present;
ii - consider the time frame immediately after the corresponding position as the time frame of establishment or obsolescence.

The previously described algorithm is able to identify items classified as *established* or *obsolete* according to our defined criteria, but not items evaluated as *short-lived* — which are classified as *random* by it. In Section 5.5, we provide a case study in which we suggest a way of adapting our method for this specific situation.

In the next section, we apply our algorithm to a real corpus, supplying five case studies to illustrate its usage and some of its potential for producing interesting observations.

5 - Case studies

In order to demonstrate the applicability of the algorithm proposed in Section 4.3, we applied it to the Corpus of Historical American English (COHA). This corpus contains more than 100,000 texts from different sources (fiction and non-fiction books, magazines, and newspapers) published in the United States of America from 1810 to 2009 (Davies, 2012), and can be explored online and downloaded from its webpage[6]. In this work, we use the case-insensitive list of unique words[7] (types), annotated with part of speech (PoS) tags. This list contains the frequency of each pair (word + PoS tag) in each of the twenty decades spanned by the data. In this way, it is often possible to differentiate between homonyms (e.g. *light*, that can be tagged as adjective, noun, verb and others). We also removed all words classified with the tags for ‘formula’, ‘proper noun’ (neutral for number, singular and plural), ‘letter of the alphabet’ (singular and plural), ‘foreign word’ (such as *arbre*, *bueno* and *deum*) and

6 https://corpus.byu.edu/coha/

7 Here, we define a *word* simply as a string of characters uninterrupted by a space. It deserves mentioning that the downloadable COHA frequency data excludes words that occur less than three times in total in the corpus.

'unclassified word' (which includes ideophones like *bang-bang*, unrecognizable words such as *carige*, exclamations like *gotcha* and recognizable words whose context is apparently unexpected). In total, we analyze 381,698 pairs of word + PoS tag in this corpus. As mentioned in Section 4.1, in these case studies we set the boundary between a `0` and a `1` in a really low relative frequency (0,00000001% of the corpus size) — so, the mere presence of a word in a time frame is enough to assign a `1` to it. By using this straightforward criterion, our goal is to show that even a method based on the simple presence/absence of items in specific time frame is able to rapidly bring interesting and useful results.

Having selected the corpus to work with, we need to decide on the value of *T*, i.e., the maximum threshold for how many deviations from the perfect patterns we can accept so to advocate for the establishment or the obsolescence of the analyzed items. Although, as mentioned in Section 4.3, the decision must to some degree be *ad hoc*, it should at least be backed up by an explicit criterion. Our approach here is to look at the statistics of establishment of words using the perfect pattern (no deviations) as a baseline: if the number of words that get established in different decades allowing for *d* amount of deviations is consistently proportional to the number of words that get established under the zero-deviation criterion, then the given value of *d* is acceptable. But how should 'consistently proportional' be defined? Here, we look at the time series for the proportion of words that became established in each decade out of all words in the decade using different values of $d > 0$, and correlate these numbers with the corresponding numbers for the zero-deviation curve. If the *p*-value of a Pearson correlation is below 0.05 for a given value of *d*, then that amount of deviation is taken to be acceptable. In our case, it is only for $d = 1$ that we find an acceptable correlation: $p = 0.0047$, $\rho = 0.605$; for $d = 2$ we already get $p = 0.0569$ and the correlation goes down to $\rho = 0.4323$. Results continue to get worse as more deviations are allowed for. Thus, it is clear that too much noise would be admitted into any statistics on the establishment of new words (and presumably on their obsolescence as well) if more than one deviation is considered acceptable in this case. For one deviation, the observations will also contain some noise, but more (and still reliable) data will be included[8].

As an illustration of how a few words are evaluated by our algorithm in COHA, Table 2 displays the outcomes of attempts to detect established/obsolete words using respectively the naive (zero-deviation criterion) approach and our proposed method implementing the one-deviation criterion. The words selected for illustration are all singular common nouns present in COHA. Words are marked with (a) when they represent cases in which their first/last attestation matches the outcomes of both algorithms; with (b) when the naive rules cannot determine their date of establishment/obsolescence and our algorithm finds that the first or last occurrences are, respectively, also the decades of establishment or obsolescence; with (c) when the naive rules again cannot tell their date of establishment/obsolescence and our algorithm now finds that the first or last occurrences are, respectively, *not* the decades of establishment and obsolescence; with (d) when the decade of establishment/obsolescence is considered random by both methods. The (b) and (c) cases are particularly relevant since they illustrate data that would be lost from the purview of a study of lexical establishment or

8 We stress that the decision on the value of this maximum threshold of deviations allowed must to some degree be *ad hoc*: since there are no 'right' and 'wrong' sets of established/obsolete items, this threshold depends on whether the researcher desires to obtain more comprehensive lists or more restricted ones — for the former, a higher threshold could be stipulated; for the latter, a lower value should be set. The point about using correlations and the *p*-value is that the distribution with one deviation from the perfect pattern is significantly similar to a distribution without any deviation, so the deviation can arguably be ignored.

obsolescence if no deviations were admitted.

Table 2. Outcomes of attempts to detect established/obsolete words using a first/last attestation approach, an algorithm following naive rules and the proposed algorithm (with a one-deviation criterion) in a selection of words present in the Corpus of Historical American English (COHA). Each time frame represents a decade, ranging from 1810s to 2000s. The examples chosen are all words tagged as 'singular common noun'.

Word	**Observed diachronic sequence**	**First attestation**	**Outcome according to**	
			naive rules	**proposed algorithm**
(a) *victrola*	00000000001111111111	1910s	established (1910s)	established (1910s)
(b) *snatcher*	00000000110111111111	1890s	random	established (1890s)
(c) *bulldozer*	00000001000001111111	1880s	random	established (1940s)
(d) *wife-murderer*	00001011101100010010	1850s	random	random
		Last attestation		
(a) *secresy*	11111111111100000000	1920s	obsolete (1930s)	obsolete (1930s)
(b) *gratulation*	11111111111111010000	1960s	random	obsolete (1970s)
(c) *destraction*	11100001000000000000	1880s	random	obsolete (1840s)
(d) *unfeelingness*	00110000110110100100	1980s	random	random

5.1 - Case 1: Statistics on established and obsolete words

Figure 1 shows the percentage of words that became established (left figure) and obsolete (right figure) per decade in COHA according to our algorithm and using the one-deviation criterion. In the left figure, the U-shaped nature of the curve concerning the establishment of words considering the whole corpus is easily explained by two factors that must always be acknowledged by the researcher: first, the proportion of words that had not appeared previously in the corpus is necessarily higher in the first time frames than in the next ones, as a consequence of the phenomenon known as Herdan's or Heaps' law (Herdan, 1964; Heaps, 1978), according to which vocabulary size grows slowly compared to the size of the document/corpus; second, the proportion of words arisen in a certain decade that are consistently present in the following ones (i.e., the words considered established conforming to our criteria) is necessarily higher in the last time frames than in the previous ones, because most of these recently established words did not have time to become obsolete yet. To demonstrate these two effects more precisely, we include two additional curves in the graph, corresponding to the percentage of words that became established in a given decade considering only certain time windows (six- and eight-decade windows). In other words, we reduce the corpus to sliding windows of six and eight decades in order to decrease the 'advantage' that early and late decades hold compared to middle decades. For these two additional curves, however, we are employing a zero-deviation criterion, since one deviation in a universe of only a few decades might be considered disproportionate. These additional curves do not display such a clear U-shaped nature, even though the one regarding six-decade windows still slightly reflects this pattern especially in its left tail. Also, both exhibit the same shape, suggesting that the proportion of established words among all words in a given decade is similar across time and that the use of different windows in this case might be no more than a question of how much data one wishes to consider: around 3% for six-decade windows vs. around 2% for eight-decade windows.

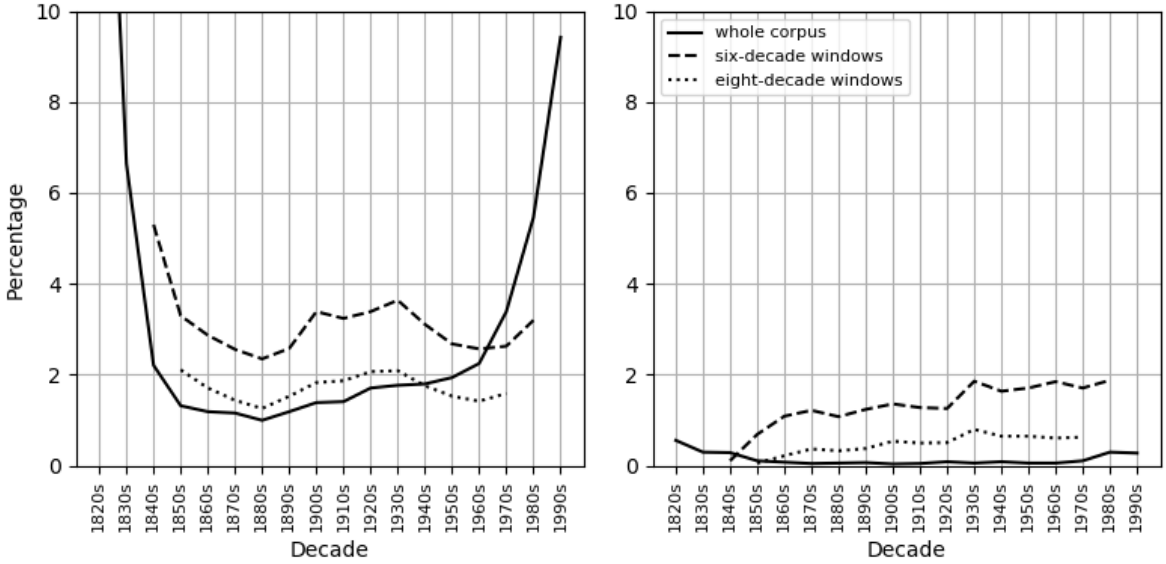

Figure 1. Percentage of words that became established (left) and obsolete (right) per decade using a one-deviation criterion and applying six- and eight-decade windows combined with a zero-deviation criterion. Curves comprise different time spans according to the sizes of the sliding windows.

Regarding the right figure, we observe that the proportion of words that became obsolete among all words in a particular decade is also more or less constant, with lower percentages than the ones referring to established words. Additional research must be carried out to more precisely understand the meaning of these results and their implications for language dynamics.

5.2 - Case 2: Characteristics of established and obsolete words

Investigating certain characteristics of words considered established or obsolete according to our proposed algorithm is also a possible line of study. Figure 2, for example, shows the average length (in number of characters) of the words that became established and obsolete in given decades. Here, we observe an irregular shape of the curve concerning words that became obsolete, but a consistently positive slope in the curve regarding established words, indicating a persistent increase in the average length of words established in the corpus across time — that goes from around eight and a half characters in the mid-19th century to almost ten characters in the second half of the 20th century. Since COHA is balanced by genre across time (Davies, 2012), this finding should in principle not be attributed to artifacts of the corpus (such as a potential increase in the proportion of scientific literature, for example). Additional investigation should be conducted to better understand this phenomenon, presumably employing other data and associating this results with the abundant previous work on word length (Grzybek, 2007).

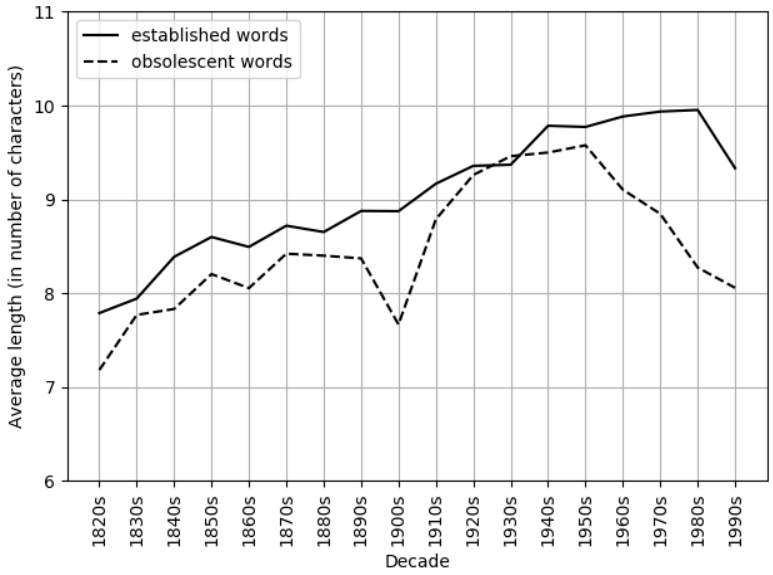

Figure 2. Average length (in number of characters) of words that became established and obsolete in given decades using a one-deviation criterion.

Different analyses can be carried out also considering the PoS tags of the words in the corpus. Figure 3 depicts the percentage of parts of speech (grouped as ‘adjective’, ‘adverb’, ‘noun’, ‘verb’ and ‘other’) among words that became established (left figure) and obsolete (right figure) in each decade. Among the established words, we visually notice a descending trend in the proportion of verbs and an ascending trend in the proportion of adjectives across the decades. The other curves are not consistently rising or falling — although, if we consider only the time period starting in the 1960s, we do observe a tendency for the proportion of nouns among the established words to increase. Regarding the words that became obsolete, the curve that represents nouns seems to exhibit a downward trend, while the others show constant fluctuation through time. Again, the fact that COHA is balanced by genre across time suggests that these patterns, in principle, should not be due to artifacts of the corpus, even though additional investigation is needed to better comprehend the phenomena reported here.

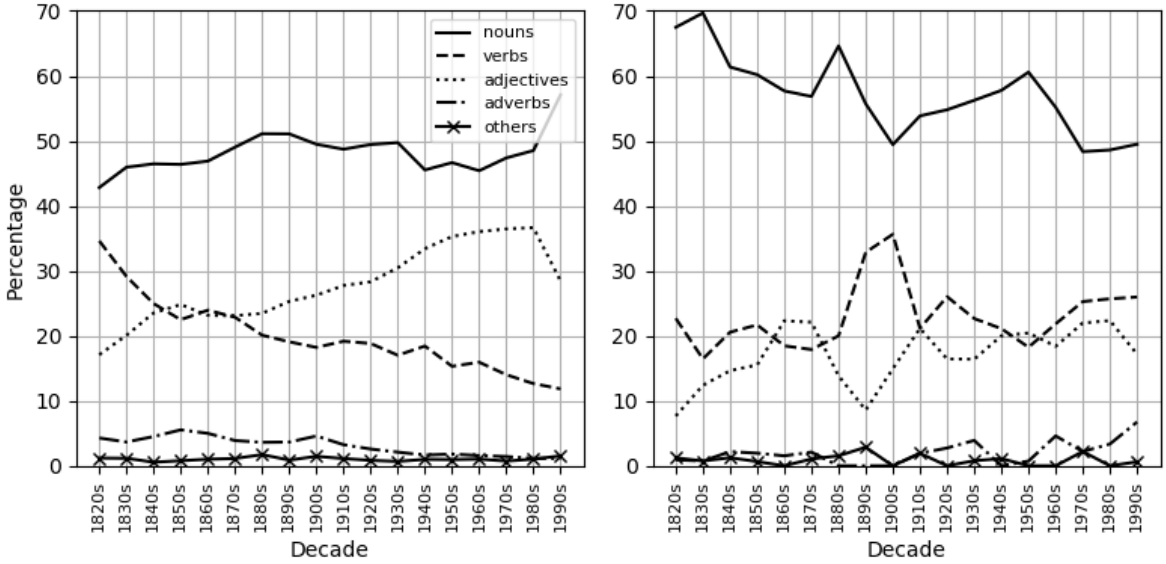

Figure 3. Percentage of different parts of speech among words that became established (left) and obsolete (right) in different decades.

5.3 - Case 3: Lexical heritage from past decades

The lexicon of every language at time *t* embodies strata from different periods in time during which new words that are still used at *t* became established. We are now onto a bit of 'stratigraphy', employing our algorithm to generate lists of those words established in different decades that are today the most popular in the corpus. More precisely, we select, from the words established in each decade between the 1850s and the 1980s, the fifty words that are most frequent in the 2000s. This ensures that we capture a portrait of today's lexical heritage from past decades which is both reasonably detailed and still salient to speakers of American English.

The result of the selection procedure is displayed in Table 3. After grouping these words into semantic categories (e.g. by using tools like Empath (Fast et al., 2016) or LIWC (Tausczik and Pennebaker, 2010)) or building networks (e.g. by a co-occurrence metric), it would be possible to make some generalizations concerning which semantic domains have been major contributors to these different historical strata or to determine the overall relationship among words established in a given decade[9].

Impressionistically, for instance, it seems that the 1870s gave us much vocabulary relating to the built environment, such as *hallway*, *downtown*, *driveway*, *taxi*, *headlights* and *neon*. The 1880s were big on sports, cf. *golf*, *hockey*, *olympics*, *coaching*, *scoring*. The 1890s were innovative in the communication domain, see *movie*, *television*, *wireless*, *phones*. The 1910s opened the written language to include words that would have been considered too obscene to print earlier: *fuck*, *goddam*, *dick*. The 1920s introduced several relatively abstract concepts relating to workflow: *coordinator*, *feedback*, *processing*, *implementation*, *operational*. The

9 These generalizations, however, should be made carefully. Even if the corpus is balanced by genre across time (which is the case of COHA), the topics covered by the texts themselves might vary systematically (and not nicely randomly) over time. A possible way of mitigating this (potential) issue could be to implement a topic detection method, such as latent Dirichlet allocation (LDA) (Blei et al., 2003), in order to ensure that topics are coherent over time.

list of the 2000s most frequent words stemming from the 1940s does not reveal that a big war happened; instead, for instance, we see elements of such an everyday affair as food consumption: *supermarket*, *microwave*, *fridge*, *burgers*, *yogurt*. The 1950s show nascent environmental concerns: *pesticides*, *recycling*, *environmentally*, *pollutants*. Apart from giving us the concept of *lifestyle(s)* in general, the 1970s also showed news in different domains of lifestyle, such as the food domain, cf. *tofu*, *fast-food*, *sushi*, *veggies*. During the 1980s, the (personal) computer is the single most dominant factor in lexical innovation: *laptop(s)*, *database(s)*, *pcs*, *algorithms*, *download*, *firewall*. As we move closer to the present, it is predictable that some of the lexical legacy encountered is going to be short-lived, and some of the terms from the 1970s and 1980s, indeed, already feel somewhat outdated.

A comprehensive study of the lexical legacy of different periods in current American English could probably use a larger selection and, as mentioned, systematic methods of defining semantic domains and networks. In any case, we show that our algorithm is effective for identifying the lexical material needed for such research.

Table 3. Lists of common words (+ PoS tags) in decade 2000s that were established in the corpus in a particular previous decade. Words are ordered according to their frequency in decade 2000s. The PoS tags that occur in this and in the following tables are[10]: *cc*: coordinating conjunction; *cs*: subordinating conjunction; *ii32*: general preposition (as part of a sequence); *jj*: general adjective; *jjr*: general comparative adjective; *mc*: cardinal number, neutral for number; *mc2*: plural cardinal number; *nn*: common noun, neutral for number; *nn1*: singular common noun; *nn122*: singular common noun (as part of a sequence); *nn2*: plural common noun; *nna*: following noun of title; *nno*: noun, neutral for number; (continues)

1850s	1860s	1870s	1880s	1890s	1900s	1910s
scientists_nn2	photo_nn1	phone_nn1	radio_nn1	movie_nn1	computer_nn1	gon_vvgk
experts_nn2	baseball_nn1	programs_nn2	skills_nn2	na_to	global_jj	electronic_jj
bike_nn1	photograph_vv0	photos_nn2	researchers_nn2	television_nn1	shit_nn1	aids_nn1
regional_jj	cigarette_nn1	hallway_nn1	golf_nn1	environmental_jj	movies_nn2	servings_nn2
detective_nn1	options_nn2	long-term_jj	parking_nn1	nuclear_jj	weekend_nnt1	sox_nn2
focused_vvn	cops_nn2	makeup_nn1	ceo_nn1	basketball_nn1	soccer_nn1	pickup_nn1
tablespoons_nn2	typically_rr	downtown_jj	soviet_jj	garage_nn1	calories_nnu2	agenda_nn1
strategies_nn2	protein_nn1	focus_vvi	networks_nn2	ta_to	jazz_nn1	helicopter_nn1
terrorists_nn2	telephone_nn1	classroom_nn1	techniques_nn2	overall_jj	coverage_nn1	fuck_nn1
users_nn2	concept_nn1	diabetes_nn1	ratings_nn2	therapy_nn1	genes_nn2	kidding_vvg
shorts_nn2	dna_nn1	bacteria_nn2	nonetheless_rr	terrorist_jj	muslim_jj	featuring_vvg
scientist_nn1	genetic_jj	driveway_nn1	korean_jj	clinic_nn1	aircraft_nn	lipstick_nn1
technique_nn1	ethnic_jj	racial_jj	consultant_nn1	mommy_nn1	someday_rt	goddamn_jj
ongoing_jj	grabs_vvz	immune_jj	hockey_nn1	wireless_nn1	quarterback_nn1	teammates_nn2
strategic_jj	backyard_nn1	colorful_jj	viewers_nn2	gym_nn1	technologies_nn2	windshield_nn1
grill_nn1	cigarettes_nn2	taxi_nn1	olympics_nn2	islamic_jj	pm_ra	someplace_rl
aluminum_nn1	cd_nn1	scenario_nn1	residential_jj	phones_nn2	rookie_nn1	lineup_nn1
parked_vvn	concepts_nn2	cultures_nn2	toxic_jj	basically_rr	emissions_nn2	parked_vvd
fake_jj	spectacular_jj	palestinian_jj	subway_nn1	scheduled_vvn	suitcase_nn1	nationwide_rl
downtown_rl	terrain_nn1	diagnosed_vvn	meaningful_jj	awareness_nn1	buddy_nn1	part-time_jj
focusing_vvg	focused_vvd	pinta_nn1	sneakers_nn2	initially_rr	activists_nn2	cafeteria_nn1
photographer_nn1	specialist_nn1	gasoline_nn1	predictable_jj	analysts_nn2	hispanic_jj	backseat_nn1
treatments_nn2	shotgun_nn1	adolescents_nn2	championships_nn2	tablespoon_nn1	muslims_nn2	prestigious_jj
trailer_nn1	sensed_vvd	evaluation_nn1	binoculars_nn2	sector_nn1	regulatory_jj	helicopters_nn2
bicycle_nn1	worldwide_rl	victorian_jj	foyer_nn1	sweater_nn1	expertise_nn1	dick_nn1
flipped_vvd	fingertips_nn2	semester_nn1	asset_nn1	coastal_jj	airplane_nn1	motivated_vvn
gestured_vvd	aging_jj	bartender_nn1	dioxide_nn1	full-time_jj	jeep_nn1	vitamins_nn2
canyon_nn1	underwear_nn1	halloween_nnt1	initiatives_nn2	locals_nn2	routinely_rr	sponsored_vvn
biology_nn1	clarity_nn1	cardboard_jj	coaching_nn1	homework_nn1	diesel_nn1	podium_nn1
ambulance_nn1	optimistic_jj	collaboration_nn1	heck_nn1	flashlight_nn1	skiing_nn1	limo_nn1
variables_nn2	technological_jj	headlights_nn2	orientation_nn1	weekends_nnt2	grid_nn1	overseas_jj
institutional_jj	emotionally_rr	fingernails_nn2	overseas_rl	homeland_nn1	researcher_nn1	recordings_nn2
slowed_vvd	yanked_vvd	housing_vvg	focuses_vvz	ok_jj	minimal_jj	airplanes_nn2
buses_nn2	specialists_nn2	starters_nn2	kilometers_nnu2	behaviors_nn2	feminist_jj	consultants_nn2
providers_nn2	obsession_nn1	neon_nn1	interactive_jj	hiking_vvg	robots_nn2	postwar_jj
productivity_nn1	starter_nn1	ramp_nn1	penis_nn1	sexually_rr	syndrome_nn1	planners_nn2
businessman_nn1	headlines_nn2	output_nn1	unpredictable_jj	bombing_nn1	carbohydrate_nn1	viruses_nn2
interactions_nn2	lethal_jj	interviewed_vvn	technician_nn1	deadline_nn1	priorities_nn2	lightweight_jj
cosmic_jj	zoo_nn1	overweight_jj	touchdown_nn1	motel_nn1	all-star_jj	protesters_nn2
overly_rr	raiders_nn2	plasma_nn1	seasonal_jj	motivation_nn1	coconut_nn1	touchdowns_nn2
vibrant_jj	motors_nn2	shortage_nn1	shack_nn1	unemployment_nn1	prostate_nn1	flips_vvz
ironic_jj	peanut_nn1	microphone_nn1	capitalism_nn1	hike_nn1	nutrients_nn2	artwork_nn1
livestock_nn	innings_nn	format_nn1	hmm_uh	trauma_nn1	buddies_nn2	campuses_nn2
yep_uh	broccoli_nn1	developers_nn2	arthritis_nn1	catalog_nn1	containers_nn2	cleanup_nn1
heartbeat_nn1	awesome_jj	investor_nn1	sweetie_nn1	motorcycle_nn1	entrepreneurs_nn2	paranoid_jj
armored_jj	fictional_jj	outcomes_nn2	entrepreneur_nn1	kinda_rr	trillion_nno	highlight_nn1
detectives_nn2	aftermath_nn1	touch_ii32	abs_jj	payroll_nn1	hormones_nn2	breathtaking_jj
toddler_nn1	canned_jj	finals_nn2	scoring_nn1	spotlight_nn1	artifacts_nn2	aspirin_nn
eerie_jj	evolutionary_jj	developer_nn1	backdrop_nn1	vodka_nn1	featured_vvd	small-town_jj
inning_nn1	vulnerability_nn1	biologist_nn1	comeback_nn1	short-term_jj	psychiatrist_nn1	comics_nn2

10 COHA employs PoS tags from the UCREL CLAWS7 Tagset. The complete description of these tags, including examples, is available at http://ucrel.lancs.ac.uk/claws7tags.html

Table 3. (continued) *nnt1*: temporal noun, singular; *nnt2*: temporal noun, plural; *nnu*: unit of measurement, neutral for number; *nnu2*: plural unit of measurement; *ra*: adverb, after nominal head; *rl*: locative adverb; *rr*: general adverb; *rt*: quasi-nominal adverb of time; *to*: infinitive marker; *uh*: interjection; *vv0*: base form of lexical verb; *vvd*: past tense of lexical verb; *vvg*: -ing participle of lexical verb; *vvgk*: -ing participle catenative; *vvi*: infinitive; *vvn*: past participle of lexical verb; *vvz*: -s form of lexical verb.

1920s	1930s	1940s	1950s	1960s	1970s	1980s
okay_rr	okay_jj	sidebar_nn1	backpack_nn1	affordable_jj	online_jj	headnote_nn1
video_nn1	computers_nn2	girlfriend_nn1	infrastructure_nn1	mets_nn2	lifestyle_nn1	high-tech_jj
iraqi_jj	pizza_nn1	t-shirt_nn1	fuckin_rr	sustainable_jj	african-american_jj	laptop_nn1
airport_nn1	fucking_jj	online_rr	ncaa_nn1	ecosystems_nn2	suv_nn1	videos_nn2
boyfriend_nn1	israeli_jj	pc_nn1	backup_nn1	upscale_nn1	parenting_nn1	globalization_nn1
sexy_jj	wildlife_nn1	teenage_jj	freeway_nn1	medium-high_jj	genome_nn1	database_nn1
robot_nn1	fuck_vv0	teenager_nn1	t-shirts_nn2	activism_nn1	high-profile_jj	booker_nn1
cholesterol_nn1	nba_nn1	teenagers_nn2	ponytail_nn1	arguably_rr	workforce_nn1	pcs_nn2
workout_nn1	sunglasses_nn2	mainstream_jj	girlfriends_nn2	healthcare_nn1	preheat_vv0	cilantro_nn1
bikes_nn2	guidelines_nn2	radar_nn1	salsa_nn1	mantra_nn1	ceos_nn2	african-americans_nn2
airlines_nn2	siblings_nn2	upcoming_jj	linebacker_nn1	ecosystem_nn1	nonstick_nn1	hip-hop_jj
antibiotics_nn2	reportedly_rr	supermarket_nn1	spokeswoman_nn1	rehab_nn1	sitcom_nn1	high-end_jj
and/or_cc	laser_nn1	innovative_jj	sunni_nn1	trendy_jj	jihad_nn1	ppg_nnu
nonprofit_jj	playoff_nn1	asshole_nn1	steroids_nn2	wetlands_nn2	tsp_nnu	sustainability_nn1
vitamin_nn1	electronics_nn1	workplace_nn1	excerpted_vvd	lasers_nn2	dumpster_nn1	phd_nna
hometown_nn1	therapist_nn1	microwave_nn1	pakistani_jj	filmmakers_nn2	recycled_jj	condos_nn2
activist_nn1	bullshit_nn1	palestinians_nn2	robotic_jj	filmmaker_nn1	networking_nn1	counterterrorism_nn1
airline_nn1	racism_nn1	fridge_nn1	pesticides_nn2	autism_nn1	pricey_jj	rapper_nn1
briefcase_nn1	mph_nnu	stereo_nn1	superstar_nn1	hosted_vvd	carb_nn1	databases_nn2
playoffs_nn2	paperwork_nn1	supportive_jj	hosting_vvg	videotape_nn1	handheld_jj	gdp_nn1
wheelchair_nn1	processor_nn1	desktop_nn1	interface_nn1	disco_nn1	world-class_jj	biotech_nn1
yoga_nn1	fucking_rr	postseason_nn1	parameters_nn2	gays_nn2	tofu_nn1	state-of-the-art_jj
allegedly_rr	programming_nn1	sensors_nn2	spacecraft_nn	makeover_vv0	sunscreen_nn1	catwoman_nn1
coordinator_nn1	labs_nn2	monitor_vvi	recycling_nn1	multicultural_jj	deregulation_nn1	mid-level_jj
footage_nn1	tourism_nn1	israelis_nn2	award-winning_jj	broadband_jj	fast-food_jj	same-sex_jj
insulin_nn1	cds_nn2	iraqis_nn1	offseason_nn1	marina_nn1	gurney_nn1	buyout_nn1
columnist_nn1	medications_nn2	basics_nn2	surfing_vvg	attendees_nn2	mid-1990s_mc2	algorithms_nn2
creativity_nn1	gop_nn1	breakthrough_nn1	venues_nn2	hosted_vvn	ethnicity_nn1	pesto_nn1
feedback_nn1	demographic_jj	fda_nn1	nightstand_nn1	boutiques_nn2	updates_nn2	biotechnology_nn1
processing_nn1	seafood_nn1	irs_nn1	environmentally_rr	one-on-one_mc	cardio_nn1	preservice_nn1
c'm_vv0	condo_nn1	estrogen_nn1	automated_jj	entitlement_nn1	shiites_nn2	glynnis_nn1
ecological_jj	rearview_nn1	predators_nn2	high-risk_jj	benchmark_nn1	countertop_nn1	download_vvi
hormone_nn1	saudi_jj	vietnamese_jj	antioxidants_nn2	scheer_vv0	sushi_nn2	mentoring_vvg
onstage_jj	integrator_nn1	monitoring_vvg	charismatic_jj	hologram_nn1	magisterium_nn1	tugger_nn1
firefighters_nn2	behavioral_jj	bikini_nn1	oregano_nn1	fucked_vvn	transnational_jj	laptops_nn2
pipeline_nn1	low-income_jj	monitoring_nn1	nascar_nn1	debuted_vvd	cloning_nn1	stakeholders_nn2
implementation_nn1	adrenaline_nn1	airborne_jj	life-threatening_jj	real-time_jj	groundwater_nn1	sweatpants_nn2
highlights_nn2	bam_vv0	cardiovascular_jj	on-site_jj	dismissive_jj	spokesperson_nn1	wastewater_nn1
nazi_jj	condom_nn1	antibiotic_nn1	geek_nn1	fucked_vvd	trailhead_nn1	starbucks_nn2
audio_jj	hp_nn1	graffiti_nn	surreal_jj	restructuring_nn1	antidepressants_nn2	wetland_nn1
airports_nn2	homeowners_nn2	staffers_nn2	upbeat_jj	on-line_jj	hispanics_nn2	minivan_nn1
operational_jj	commercials_nn2	health-care_nn1	module_nn1	postmodern_jj	meds_nn2	pager_nn1
slacks_nn2	vinyl_nn1	late-night_jj	weaponry_nn1	real-world_jj	bikers_nn2	gawain_nn1
stairwell_nn1	plywood_nn1	thai_jj	chiles_nn2	cuz_nn1	mid-1980s_nn2	basher_nn1
nylon_nn1	dj_nn1	viewer_nn1	pollutants_nn2	lasagna_nn1	biking_vvg	creationism_nn1
input_nn1	marijuana_nn1	burgers_nn2	addictive_jj	simulations_nn2	recycling_vvg	feel-good_jj
aerobic_jj	implemented_vvn	predator_nn1	aerospace_nn1	deco_nn122	veggies_nn2	firewall_nn1
teammate_nn1	boutique_nn1	yogurt_nn1	clueless_jj	cornerback_nn1	midlife_nn1	ipo_nn1
receptionist_nn1	skiers_nn2	targeting_vvg	retro_jj	surfers_nn2	superhero_nn1	camcorder_nn1
workouts_nn2	condoms_nn2	paperback_nn1	multimedia_nn	batmobile_nn1	lifestyles_nn2	downtime_nnt1

5.4 - Case 4: Lost words

Another use of our proposed method is to generate lists of previously popular items that became obsolete in the corpus during its time span. This is interesting because, unlike innovations, obsolete items are not commonly covered in existing literature (Tichý, 2018). Here, we select, from the words that became obsolete in each decade between the 1850s and the 1980s, the ten with the highest frequency before their obsolescence. In this fashion, we display a vocabulary that was particularly relevant in the past, but that has lost terrain in

American English after some decades.

The lists of the once common words that became obsolete in COHA in a particular decade are shown in Table 4. In the second half of the 19th century, it is possible to encounter typos and spelling mistakes that ceased to appear in the 20th century (maybe partially due to the development of more accurate typing and printing techniques), such as *had'nt* (1870s), *do'nt* (1880s), *hav'nt* (1890s), *was'nt* (1890s) and *did'nt* (1890s). There are also several other words that are still easily recognizable but have obsolete or semi-obsolete spellings, including *errour* (1850s), *pennyless* (1870s), *musquitoes* (1880s), *negociation* (1890s), *villany* (1930s), *reconnoissance* (1940s), *trowsers* (1950s), and *persistency* (1960s), to name a few. In some cases, the obsolete spelling is more faithful to the etymology of the word, as in *holydays* (1880s), which became 'holidays', and *cocoa-nut* (1930s), which became 'coconut'. Further, the lists exhibit spellings that are still present in the corpus, but no longer with a specific syntactic function, such as *under* as a comparative adjective (1900s), *itself* as a singular common noun (1900s) and *notwithstanding* as a subordinating conjunction (1970s).

Table 4. Lists of common words (+ PoS tags) in previous decades that became obsolete in the corpus in a particular decade. Words are ordered according to their frequency before their obsolescence. When the word ranked in the eleventh position has the same frequency as the one in the tenth position, we include both. The meaning of each PoS tag is explained in the caption of Table 3.

1850s	1860s	1870s	1880s	1890s	1900s	1910s
errour_nn1	copy-right_nn1	had'nt_vv0	phrensy_nn1	merchandize_vv0	shakspeare_vv0	shakspeare_nn1
scymetar_nn1	hazle_nn1	ancle_nn1	sassacus_nn1	rivalship_nn1	immoveable_jj	eend_nn1
almanzor_nn1	do'st_vv0	pennyless_jj	ancles_nn2	hav'nt_vv0	under_jjr	did'st_vv0
pedrillo_nn1	pannels_nn2	wo-begone_jj	cotemporary_jj	had'st_vv0	say'st_vv0	creatur_nn1
musquetry_nn1	phrensied_jj	inartificial_jj	musquitoes_nn2	guarantied_vvn	xve_nn1	deth_vvz
inquietudes_nn2	choaked_vvd	wrapp_nn1	afford_nn1	intenseness_nn1	itself_nn1	piano-forte_nn1
renegado_nn1	fire-side_jj	teaze_vvi	holydays_nn2	was'nt_vv0	wall-street_nn1	saidst_vv0
errours_nn2	barb'rous_jj	rivalships_nn2	gallopped_vvd	negociation_nn1	pedee_nn1	thou'st_nn1
zegri_nn1	famish_jj	a'nt_vv0	apalachian_jj	cretur_nn1	xve_vv0	applauses_nn2
broad-street_nn1	fann_nn1	returnless_jj	do'nt_vv0	did'nt_nn1	sdeath_nn1	knitting-work_nn1
potawatamies_nn2	incommunicative_jj		vanquish_jj			see'st_vv0
1920s	**1930s**	**1940s**	**1950s**	**1960s**	**1970s**	**1980s**
the_nnt1	villany_nn1	csar_nn1	trowsers_nn2	acquirements_nn2	sich_vv0	intrusted_vvn
desponding_jj	prison-house_nn1	new-comer_nn1	school-master_nn1	inclosure_nn1	ball-room_nn1	arm-chair_nn1
flag-staff_nn1	nuther_vv0	custom-house_nn1	mantel-piece_nn1	persistency_nn1	now-a-days_rt	fellow-men_nn2
sportively_rr	wofully_rr	custom-house_jj	despatch_vvi	state-room_nn1	frying-pan_nn1	quitted_vvd
befel_vv0	unbiassed_jj	bethink_vv0	hill-top_nn1	upon_nn1	notwithstanding_cs	with_nn1
stopt_vv0	dew-drops_nn2	reconnoissance_nn1	aliment_nn1	intrenchments_nn2	hesitating_jj	common-place_jj
discomposed_vvn	cocoa-nut_jj	hill-tops_nn2	corner-stone_nn1	snuff-box_nn1	reprobation_nn1	fitly_rr
enginery_nn1	log-house_nn1	prayer-meetings_nn2	leipsic_nn1	strifes_nn2	banditti_nn2	unwearied_jj
school-fellows_nn2	can'st_vv0	school-boys_nn2	exhaustless_jj	guard-house_nn1	by-gone_jj	small-pox_nn
sarvice_nn1	palm-tree_nn1	sketch-book_nn1	self-complacency_nn1	heart-strings_nn2	plighted_jj	inclosed_vvn

Of particular interest is the illustration provided by these lists of the phenomenon of the historical spelling change of English compounds. According to Shertzer (1996), '[t]he usual sequence is for the words to be written separate at first, then to become hyphenated, and finally to be written solid' (p. 109). We observe that several compounds that are nowadays usually written in a solid form are present in the corpus as hyphenated compounds and that these became obsolete at some point — probably around the time when their corresponding solid form were gaining popularity. This is the case of *copy-right* (1860s), *fire-side* (1860s), *wo-begone* (1870s) (now most commonly written 'woebegone'), *piano-forte* (1910s) (now

mostly encountered as just 'piano'), *flag-staff* (1920s), *school-fellows* (1920s), *dew-drops* (1930s), *new-comer* (1940s), *corner-stone* (1950s), *state-room* (1960s), *ball-room* (1970s), *now-a-days* (1970s), *arm-chair* (1980s), *common-place* (1980s) and various others that can be recognized in the table. Nonetheless, a few compounds, such as *wall-street* (1900s) and *knitting-work* (1910s), seem to have taken the opposite direction, now being more commonly written as separate words. A comprehensive study aiming to analyze this phenomenon in a quantitative fashion could benefit from our proposed method to obtain these lists of obsolete items per time frame and investigate how different factors (e.g. time, accumulated frequency, sudden frequency rise/fall) act and impact this process of orthographic variation and change.

5.5 - Case 5: Short-lived words

The method described in Section 4.3 is able to assist in the identification of items classified as *established* or *obsolete*, but not of items evaluated as *short-lived*. Here, we provide a short case study in which we suggest a way of adapting it for this specific purpose. Our goal is to find words that flared up in the corpus for some time and then, still during the period covered by the corpus, disappeared. According to our previously mentioned criteria, these words are considered neither established (since they are already gone) nor obsolete (since they are not part of the corpus in its initial period), but in some cases it might be interesting to analyze them in order to investigate the process of lexical variation and change in more detail.

A possible way of adapting our method to the case of short-lived items is by applying the proposed algorithm to selected intermediate subcorpora. One solution would be to look for items whose diachronic sequences hold only `0`s in their extreme time frames, such as in `0001111000`, then cut off the extremes of the corpus (say, the n_1 time frames in the beginning and the n_2 time frames in the end of the time span covered by the corpus) and, finally, apply the algorithm only to the remaining intermediate sequences, looking for established, obsolete and permanent items in these subcorpora.

For the present exploratory purposes we adapted our method to handle cases of words that did not appear in COHA before the 1860s and disappeared again no later than the 1950s — in other words, these items are present neither in the five first nor in the five last time frames of the corpus ($n_1 = n_2 = 5$). We then applied our algorithm considering just this subsection of the corpus. We extracted words evaluated as *permanent* — which are, of course, perfect cases of short-lived words, presenting the diachronic sequence `[00000]1111111111[00000]`[11]. We also gathered other not-so-short-lived words evaluated as *established* and *obsolete* in the subsection of the corpus, but only those that appeared in at least eight decades and with no deviations allowed[12].

The words that emerged from this analysis are listed alphabetically in Table 5. The vast

11 The `0`s in between square brackets correspond to the extremes of the corpus that were cut off.

12 That is, those which, for the period of the subcorpus studied, presented the diachronic sequences `01111111111`, `1111111110`, `0011111111` and `1111111100`.

majority of them are compounds (either hyphenated or solid), short-lived spelling variants and bona fide words that came and went. Among the hyphenated compounds, we find words such as *farm-lands*, *hair-pin* and *saddle-bag* — all of them more commonly written in a solid form nowadays. These data are useful for the study of the historical spelling change of English compounds mentioned in Section 5.4. Words such as *comp'ny*, *yisterday* and *s'posin* are examples of short-lived spelling variants. The comparative adjective *humaner* (meaning *more humane*) and the nouns *leisureliness* (*leisurely* + *-ness*) and *stereopticon* (an old type of slide projector) are interesting examples of short-lived items found here: when searching on another source, the Google Books Ngram Viewer[13], we find that all of them exhibit a similar frequency pattern, peaking around the 1920s.

Table 5. Words (+ PoS tags) classified as *short-lived* according to the adaptation of our method and considering the period between the 1860s and the 1950s. Words are alphabetically ordered. The meaning of each PoS tag is explained in the caption of Table 3.

a-beatin_nn1	crep_nn1	ha'r_nn1	race-track_nn1
a-laughin_nn1	dilapidated-looking_jj	hair-pin_nn1	rose-petals_nn2
a-puttin_nn1	dish-towels_nn2	hay-wagon_nn1	s'posin_nn1
a-quiver_vv0	dust-heap_nn1	hereinbefore_rr	sabe_vvi
a-sittin_nn1	ear-drums_nn2	herse'f_nn1	saddle-bag_nn1
all-rail_jj	earnin_nn1	hez_vv0	spoilin_nn1
alongshore_nn1	east-bound_jj	high-tariff_jj	staff-officer_nn1
baggageman_nn1	farm-hands_nn2	humaner_jjr	station-master_nn1
bath-chair_nn1	farm-lands_nn2	ice-floe_nn1	stereopticon_nn1
bird-shot_nn1	field-glass_nn1	idealizing_jj	street-cars_nn2
black-fringed_jj	field-glasses_nn2	jumping-jack_nn1	talesmen_nn2
bodder_vvi	fitten_vvn	leisureliness_nn1	tek_vvi
bofe_nn1	food-supply_nn1	lucile_nn1	trades-union_nn1
bread-winner_nn1	foregathered_vvd	myse'f_nn1	unfoldment_nn1
broncho_nn1	forehanded_vvn	pack-train_nn1	up-train_nn1
burled_vvn	four-bit_jj	pay-rolls_nn2	w'at_nn1
catchee_nn1	full-armed_nn1	pepsin_nn1	w'en_jj
chromos_nn2	garden-party_nn1	play-actin_nn1	water-bottle_nn1
coat-sleeves_nn2	glarin_nn1	pony-cart_nn1	weazened_vvd
comp'ny_jj	groceryman_nn1	prohibitionist_jj	wedding-bells_nn2
consul-general_jj	grouped_jj	pulse-beats_nn2	yisterday_nn1

These results are just an illustration of the kind of content that can be obtained from such an analysis. It is important to notice that looking for short-lived items is not, in principle, one of the goals of the method introduced in this paper, and that the adaptation presented in this case study is just a workaround. The main pitfall of this adaptation is that it depends on the selection of specific subsections of the corpus to be analyzed by the researcher. A possible goal for future work is to design and develop a specific and more effective method for finding short-lived items in diachronic corpora.

6 - Concluding remarks

In the field of corpus linguistics, the analysis of diachronic corpora with the goal of explaining diverse phenomena in human languages is becoming increasingly widespread. In this context, we need methods and procedures aiming to discover trends and patterns in the dynamics of a language as we process big amounts of text computationally. With the present contribution, we hope to specifically generate more interest in the birth and death of

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

components such as words, expressions and grammatical constructions in corpora that span over time.

Here, we introduce the notions of *establishment* and *obsolescence* as complementary to the trivial concepts of first and last attestations of linguistic items in diachronic corpora. Subsequently, we propose an algorithm to identify the time period of establishment and obsolescence of linguistic items based on their frequency in a diachronic corpus. This algorithm may be employed for the analysis of any linguistic item, be it lexical, phonological or morphosyntactical.

We demonstrate the applicability of our proposed algorithm using a real corpus spanning 200 years of data and supplying case studies concerning the character of words that got established and obsolete in American English in different periods. Among the outcomes of these case studies is the observation that the percentage of established words among all words across decades fluctuates without showing a specific upward or downward trend. We also found that the proportion of adjectives among new words has increased steadily over the past two centuries, mostly mirrored by a decrease in the proportion of new verbs. Then, we provided a sketch study of the lexical heritage in American English, identifying words that became established in different decades and are still frequent in the 2000s. We also looked at obsolescent vocabulary — vocabulary that was previously frequent but has been getting lost over the decades. Finally, we briefly investigated whether the method could be adapted to find short-lived words — words that flared up in the corpus for some time and then disappeared. These sketch studies are mainly presented with the goal of motivating future studies employing the method presented here.

It may be obvious but still it is necessary to recall that a corpus is different from a language. As a consequence, when we consider the establishment or the obsolescence of a linguistic item in a *corpus*, we are not necessarily referring to the establishment or the obsolescence of this item in a *language*. This distinction is particularly relevant when we deal with corpora based on written texts (like COHA itself or the Google Books corpus) — since, for instance, an item might be used for a long time in the oral language before it gets established in the written register. When considering the whole language, it is clear that the algorithm can only identify the decade during or *before which* (*ante quem*) a word became established or the decade during or *after which* (*post quem*) a word became obsolete. This situation is of course due to the fact that ‘it is much simpler to prove that something exists (...) than to prove that something does not exist’ (Tichý, 2018: 82). This fact becomes even clearer if we think about the application of our method to domain-specific corpora (consisting of academic, legal, medical etc. texts): the results will of course reflect the specificity of the analyzed data.

As stated by Hilpert and Mair (2015), it is imperative to demonstrate ‘how the use of corpus data allows researchers to go beyond the mere statement that a grammatical change happened, and to address the questions of **when** and **how** something happened’ (Hilpert and Mair, 2015: 199, emphases in original). With our theoretical discussion, our proposed algorithm, and the case studies that were presented here, we hope to have taken a step in this direction.

References

Algeo, John, and Adele S. Algeo (eds.). 1993. *Fifty years Among the New Words: a dictionary of neologisms, 1941–1991*. Cambridge: Cambridge University Press.

Ayto, John. 1989. *The Longman register of new words*, vol. 1. Harlow: Longman.

Ayto, John. 1990. *The Longman register of new words*, vol. 2. Harlow: Longman.

Ayto, John. 1999. *Twentieth century words*. Oxford: Oxford University Press.

Bauer, Laurie. 2002. 'Inferring variation and change from public corpora' in J. K. Chambers, Peter Trudgill and Natalie Schilling-Estes (eds.) *The handbook of language variation and change*, pp. 97-114. Hoboken: Wiley Blackwell.

Biber, Douglas and Bethany Gray. 2011. 'Grammatical change in the noun phrase: the influence of written language use', *English Language and Linguistics* 15 (2), pp. 223–250.

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. 'Latent dirichlet allocation.' *Journal of Machine Learning Research* 3, pp. 993-1022.

Blumenthal-Dramé, Alice. 2012. *Entrenchment in usage-based theories: what corpus data do and do not reveal about the mind*, vol. 83 of Topics in English Linguistics. Berlin: De Gruyter Mouton.

Bochkarev, Vladimir, Valery Solovyev and Søren Wichmann. 2014. 'Universals versus historical contingencies in lexical evolution', *Journal of the Royal Society Interface* 11 (101).

Croft, William. 2000. *Explaining language change: an evolutionary approach*. Harlow: Pearson Longman.

Davies, Mark. 2012. 'Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English', *Corpora* 7 (2), pp. 121-157.

Fast, Ethan, Binbin Chen and Michael S. Bernstein. 2016. 'Empath: understanding topic signals in large-scale text', *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4647-4657.

Gries, Stefan Th. and Martin Hilpert. 2008. 'The identification of stages in diachronic data: variability-based neighbour clustering', *Corpora* 3 (1), pp. 59-81.

Grzybek, Peter. 2007. 'History and methodology of word length studies', in Peter Grzybek (ed.) *Contributions to the science of text and language: word length studies and related issues*, pp. 15-90. Berlin: Springer.

Hamilton, William L., Jure Leskovec and Dan Jurafsky. 2016. 'Diachronic word embeddings reveal statistical laws of semantic change', *Proceedings of the 54th Annual Meeting of the*

Association for Computational Linguistics, pp. 1489-1501.

Heaps, Harold Stanley. 1978. *Information retrieval: computational and theoretical aspects*. New York: Academic Press.

Herdan, Gustav. 1964. *Quantitative linguistics*. London: Butterworth.

Hilpert, Martin and Christian Mair. 2015. 'Grammatical change' in Douglas Biber and Randi Reppen (eds.) *The Cambridge handbook of English corpus linguistics*, pp. 180-200. Cambridge: Cambridge University Press.

Hilpert, Martin and Stefan Th. Gries. 2009. 'Assessing frequency changes in multistage diachronic corpora: applications for historical corpus linguistics and the study of language acquisition', *Literary and Linguistic Computing* 24 (4), pp. 385-401.

Hinrichs, Lars and Benedikt Szmrecsanyi. 2007. 'Recent changes in the function and frequency of Standard English genitive constructions: a multivariate analysis of tagged corpora', *English Language and Linguistics* 11 (3), pp. 437-474.

Hundt, Marianne and Christian Mair. 1999. ' "Agile" and "uptight" genres: the corpus-based approach to language change in progress', *International Journal of Corpus Linguistics* 4, pp. 221-242.

Knowles, Elizabeth and Julia Elliot (eds.). 1997. *The Oxford dictionary of new words*. Oxford: Oxford University Press.

Langacker, Ronald W.. 1987. *Foundations of cognitive grammar: theoretical prerequisites*, vol. 1. Stanford University Press.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak and Erez Lieberman Aiden. 2011. 'Quantitative analysis of culture using millions of digitized books', *Science* 331 (6014), 176-182.

Moon, Rosamund. 2010. 'What can a corpus tell us about lexis?' in Anne O'Keeffe and Michael McCarthy (eds.) *The Routledge handbook of corpus linguistics*, pp. 197-211. New York: Routledge.

Pagel, Mark, Quentin D. Atkinson and Andrew Meade. 2007. 'Frequency of word-use predicts rates of lexical evolution throughout Indo-European history', *Nature* 449, pp. 717-720.

Perc, Matjaž. 2012. 'Evolution of the most common English words and phrases over the centuries', *Journal of The Royal Society Interface* 9 (77), pp. 3323-3328.

Petersen, Alexander M., Joel Tenenbaum, Shlomo Havlin and H. Eugene Stanley. 2012.

‘Statistical laws governing fluctuations in word use from word birth to word death’, *Scientific Reports* 2 (313).

Schmid, Hans-Jörg. 2007. ‘Entrenchment, salience, and basic levels’ in Dirk Geeraerts and Hubert Cuyckens (eds.) *The Oxford handbook of cognitive linguistics*, pp. 117-138. Oxford: Oxford University Press.

Shepherd, Tania Maria Granja. 2014. ‘Changing “faces”: a case study of complex prepositions in Brazilian Portuguese’ in Tony Berber Sardinha and Telma de Lurdes São Bento Ferreira (eds.) *Working with Portuguese corpora*, pp. 69-88. Bloomsbury Academic.

Shertzer, Margaret D. 1996. *The elements of grammar*. New York: Macmillan Publishing Company.

Tausczik, Yla R. and James W. Pennebaker. 2010. ‘The psychological meaning of words: LIWC and computerized text analysis methods.’ *Journal of Language and Social Psychology* 29(1), pp. 24-54.

Tichý, Ondřej. 2018. ‘Lexical obsolescence and loss in English: 1700–2000’ in Joanna Kopaczyk and Jukka Tyrkkö (eds.) *Applications of pattern-driven methods in corpus linguistics*, pp. 81-103. Amsterdam: John Benjamins.

Tulloch, Sara. 1991. *The Oxford dictionary of new words: a popular guide to words in the news*. Oxford: Oxford University Press.

Vo, Anh Thuc and Ronald Carter. 2010. ‘What can a corpus tell us about creativity?’ in Anne O’Keeffe and Michael McCarthy (eds.) *The Routledge handbook of corpus linguistics*, pp. 302-315. New York: Routledge.

Widdowson, Henry G., 2000. ‘On the limitations of linguistics applied’, *Applied Linguistics* 21 (1), pp. 3-25.