COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 45 (2021) e12963

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12963

Words with Consistent Diachronic Usage Patterns are Learned Earlier: A Computational Analysis Using Temporally Aligned Word Embeddings

Giovanni Cassani,[a] Federico Bianchi,[b] Marco Marelli[c]

[a]*Department of Cognitive Science and Artificial Intelligence, Tilburg University*
[b]*Bocconi Institute for Data Science and Analytics, Bocconi University*
[c]*Department of Psychology, University of Milano-Bicocca*

Received 8 July 2020; received in revised form 15 February 2021; accepted 21 February 2021

Abstract

In this study, we use temporally aligned word embeddings and a large diachronic corpus of English to quantify language change in a data-driven, scalable way, which is grounded in language use. We show a unique and reliable relation between measures of language change and age of acquisition (*AoA*) while controlling for frequency, contextual diversity, concreteness, length, dominant part of speech, orthographic neighborhood density, and diachronic frequency variation. We analyze measures of language change tackling both the change in lexical representations and the change in the relation between lexical representations and the words with the most similar usage patterns, showing that they capture different aspects of language change. Our results show a unique relation between language change and *AoA*, which is stronger when considering neighborhood-level measures of language change: Words with more coherent diachronic usage patterns tend to be acquired earlier. The results support theories positing a link between ontogenetic and ethnogenetic processes in language.

Keywords: Age of acquisition; Language change; Temporally aligned word embeddings; Computational psycholinguistics

Giovanni Cassani and Federico Bianchi contributed equally to this work.

Correspondence should be sent to Giovanni Cassani, Department of Cognitive Science and Artificial Intelligence, Tilburg University, Warandelaan 2 5037AB, Tilburg, The Netherlands. Email: g.cassani@tilburguniversity.edu

1. Introduction

Languages change at many different levels (Croft, 2000), involving form, structure, and meaning, with these processes being influenced by social, cognitive, and cultural factors (Labov, 2001; Lieberman, Michel, Jackson, Tang, & Nowak, 2007; Thomason & Kaufman, 1992). According to recent theories that stress the role of cognitive rather than biological factors in language evolution, change likely serves the purpose of adapting the language to the need of the community of speakers (Smith, 2004; Steels, 2017). However, for a language to successfully serve its communicative purpose, it needs to preserve a certain stability, such that it can be passed to following generations (Kirby, 2001). In this paper, we zoom in on the link between language change and language acquisition, to investigate whether there exists a unique relation between words that are acquired earlier and words whose usage patterns remain consistent over time.

Crucially, our goal is not to use diachronic patterns to model the process by which an individual child learns some words before others (Braginsky, Yurovsky, Marchman, & Frank, 2019; Hills, Maouene, Maouene, Sheya, & Smith, 2009); after all, children are not exposed to the diachronic history of their native language during learning. Rather, we ask ourselves why some words, based on diachronic usage patterns, are the words that are learned earlier. The hypothesis is that the language system as a whole might evolve in such a way that, in the shifting tides of language, some words become the ones being typically learned earlier.[1] We intend to first check whether a relation between acquisition and changes in the diachronic usage patterns of lexical items exists. Then, we analyze whether such relation holds after controlling for the contemporary language landscape, as captured by measures of frequency, contextual diversity, concreteness, orthographic neighborhood density, and word length (all predictors that have been reported to influence age of acquisition (*AoA*; Braginsky et al., 2019; Hills et al., 2010), as well as by diachronic variations in frequency. If confirmed, this would entail that processes relating to linguistic stability in usage patterns capture something unique about what learners end up learning.

In pursuing this goal, our major contribution lies in the application of a novel method to quantify stability of usage patterns over time. We then use several measures derived using such method to analyze the relation between language acquisition and language evolution. While previous studies addressing this relation (Monaghan, 2014; Monaghan & Roberts, 2019; Vejdemo & Hörberg, 2016) relied on small sample sizes and expert annotations, limiting the scope of their results, we leverage recent advances in natural language processing (NLP) to track language behavior over time and learn linguistic representations in an unsupervised and scalable way (Bianchi, Di Carlo, Nicoli, & Palmonari, 2020; Di Carlo, Bianchi, & Palmonari, 2019).

Language change has typically been analyzed considering social and cultural factors (Labov, 2001). Recent studies, however, have analyzed which properties of words, gauged from usage patterns, influence their evolution over time (Christiansen & Chater, 2008). Pagel, Atkinson, and Meade (2007) showed that more frequent words tend to have fewer cognates; that is, different words in other languages within the same phylogenetic tree, which have a shared etymological origin. Moreover, Vejdemo and Hörberg (2016) reported that more frequent words and words with more synonyms have a higher probability of undergoing lexical

replacement, while polysemy reduces the likelihood that a word is replaced. Both studies focused on words from the Swadesh list (Swadesh, 1952) containing 200 English words that are considered to be fundamental terms in most languages. Winter, Thompson, and Urban (2014) also provided evidence that words that are likelier to be the origin of language change tend to be more frequent and polysemous, have more associations in free word association data, and are part of denser semantic networks.

Two recent studies further investigated the relation between cognitive factors and language evolution. Again relying on the Swadesh list, Monaghan (2014) documented a unique effect of *AoA* on cognate proliferation, with early acquired words correlating with lower rates of lexical evolution. Moreover, Monaghan and Roberts (2019) showed that early acquired words are less likely to be borrowed from other languages. This study also addressed issues of representativeness by analyzing several hundreds of target words from two different languages, Dutch and English, and documented that Swadesh lists (Swadesh, 1952) are not representatives of the whole lexicon and may thus yield unreliable results. The reported effect of *AoA* on cognate proliferation and likelihood of borrowing was interpreted in the light of converging evidence on the role of acquisition processes in shaping language evolution (Christiansen & Chater, 2008; MacNeilage & Davis, 2000). This effect suggests that a relation exists between ontogenetic and ethnogenetic language evolution, in that language learning during the life of an individual relates to how the language evolved over time in the community of speakers, as the object of learning at a given moment depends on its evolution over time and since the language learned by an individual contributes to shaping the language used by the community, in turn contributing to the continuing diachronic change of the language.

Importantly, the reported relation between *AoA* and language change remains when controlling for the effect of contemporary usage, gauged through word frequency. Frequency of use in the contemporary community and its effect on language change (Pagel et al., 2007) suggests a pressure to conform to other speakers in order to avoid misinterpretation of words often used in discourse (Boyd & Richerson, 1988).[2] On the contrary, the relation between *AoA* and lexical evolution has been hypothesized to relate to representational salience, higher for early acquired words (Monaghan, 2014). This notion of representational salience is justified on the basis of two phenomena. First, Juhasz (2005) documented faster processing of words and pictures with earlier learned labels than words and pictures with later learned labels in many different tasks. Second, studies with elderly people (Hodgson & Ellis, 1998), aphasic (Bradley, Davies, Parris, Su, & Weekes, 2006), and Alzheimer patients (Holmes, Jane Fitch, & Ellis, 2006) showed that early acquired words are retained longer. Representational salience, therefore, seems to be related to conceptual availability: The cognitive prioritization that these words receive because of their early acquisition plays a role in contributing to a greater stability of these same words.

Even though they provide insights about the relation between language ontogenesis and ethnogenesis, the reviewed studies present methodological shortcomings, primarily due to (i) limitations in sample size, (ii) potential biases in the sample, (iii) reliance on expert annotations and manually curated resources (e.g., thesauri, Swadesh lists, and lexica), which make these methods hard to scale, and (iv) indirect measures of language change, which abstract away from actual language use. In what follows, we discuss how these problems are addressed by our proposed operationalization of language change, which builds on recent work in NLP

on detecting semantic shifts (see Tahmasebi, Borin, & Jatowt, 2018, for a review on this research line).

While the four aforementioned limitations are distinct and could be addressed orthogonally, we propose to address them organically. Rather than using curated resources, we exploit recent NLP techniques to learn temporal word representations from diachronic corpora covering a long period of time (Bianchi et al., 2020; Bowern, 2019; Di Carlo et al., 2019; Hamilton et al., 2016b; Tahmasebi, Borin, Jatowt, & Xu, 2019). The advantages of this approach over previous solutions are several: first, it does not require any external resource, such as Swadesh lists, lexica providing estimates of lexical substitutions, and thesauri to estimate polysemy or number of synonyms. Therefore, it is more robust, more representative of general linguistic patterns, and cheaper to compute, addressing three methodological shortcomings of previous works at once. The only required resource is a sizable diachronic corpus, and such type of text data are becoming increasingly available. Given such a corpus, it is in principle possible to automatically compute measures of language change for any word in the corpus. This allows researchers to work with far larger sample sizes and to improve the representativeness of the sample as well as the robustness of observed patterns in any study targeting language change (Sagi, Kaufmann, & Clark, 2011).

This method also allows researchers to characterize language change from documented language use rather than from indirect measures, which depend on manual coding and rest on possibly questionable assumptions, addressing the fourth limitation we identified in previous studies. The idea of operationalizing language phenomena considering how languages are used by communities of speakers is well established (Firth, 1957; Harris, 1954; Wittgenstein, 1953) and has proven very useful in linguistics, psycholinguistics, cognitive science as well as artificial intelligence and NLP (Griffiths, Steyvers, & Tenenbaum, 2007; Günther, Rinaldi, & Marelli, 2019; Lenci, 2018). The specific advantage of this theoretical position in the context of the current problem and its novelty compared to previous studies on the relation between language evolution and acquisition, however, is in the possibility of actually tracking how language use changed over time and exploiting this to operationalize language change for individual lexical items (Hamilton et al., 2016a; Hilpert & Perek, 2015; Perek, 2014; Sagi et al., 2011).

In practice, we learn word representations using distributional semantics methods (Baroni & Lenci, 2010; Landauer & Dumais, 1997; Lund & Burgess, 1996; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), which showed that representations of lexical meanings can be captured via the co-occurrence patterns of words in context (Firth, 1957). For example, *dog* and *cat* will have similar lexical representations, in line with the similar way in which the two words are used in English and their similar co-occurrence patterns with other words. Within this framework, words are represented using distributed, numerical vectors embedded in a high-dimensional space (Turney & Pantel, 2010). Such distributed word representations are usually referred to as *word embeddings*, to stress the fact that a word vector representation is embedded in a particular space and can only be interpreted in relation to other representations in the same space, without the need of a priori semantic categories. Embedding spaces are geometrical spaces in which measures of proximity between word vectors can be computed, such that more similar words tend to have more similar embeddings and to be closer in space. It has been widely shown that such measures capture relevant structural and semantic

phenomena, for example, semantic associations (Caliskan, Bryson, & Narayanan, 2017; Landauer & Dumais, 1997), lexical categories (Westbury & Hollis, 2019), and valence (Hollis & Westbury, 2016). Crucially, words are closer in this high-dimensional embedding space when their overall co-occurrence patterns are more similar, rather than when simply co-occurring together often (Günther et al., 2019). Therefore, we can track differences in the co-occurrence patterns of a word over time, abstracting away from specific lexical co-occurrences.

In order to learn representations that can be used to compute language change,[3] we need diachronic corpora that are divided chronologically in different slices (or bins), such that a different word representation is learned from the language samples in each slice. In order to compare time-dependent word representations, however, it is necessary to ensure that all word embeddings exist in a vector space defined by the same coordinates (Di Carlo et al., 2019); that is, embeddings from different time slices are aligned. This passage is nontrivial: It is not enough to learn a different embedding space for each slice of a diachronic corpus because such embedding spaces are not necessarily comparable as they have been learned independently. It is therefore crucial to ensure that all embeddings exist within the same coordinate system; in our study this is achieved by first obtaining a general embedding space from the whole corpus that acts as a compass to align the embedding spaces generated from each temporal space to the main space and to each other. Once this step has been completed, relations of proximity across embeddings can thus be leveraged to quantify language change, as word embeddings reflect the usage patterns of a word in the language over time in the same reference embedding space.

Let us consider a real example to make things more concrete. Around 1920, the word *keyboard* had a very similar embedding to *violin* or *piano*, as it denoted an element of a musical instrument. Nowadays, after the technological revolution, the embedding representation for *keyboard* is more similar to those of *desktop* and *keypad*, reflecting the fact that a keyboard primarily identifies a computer part. Of course, *keyboard* still also denotes a musical instrument; however, its use changed over time, with a new meaning being introduced and becoming prevalent, thus affecting the lexical representation learned from the corpus and the relations that this representation has with the other lexical representations in the language network.

This example also highlights a further advantage of the current approach, which allows us to consider item-level properties as well as neighborhood-level properties in tracking language change (Hamilton et al., 2016a). Since time-dependent word embeddings referring to the same word exist within the same coordinate system, we can compare how similar they are, leveraging relations of proximity to quantify whether a word representation changed over time (item level). We can thus compare the embedding for *keyboard* learned from a sample of 1800 English words to the embedding for *keyboard* computed using contemporary English and measure their proximity: The farther they are, the more the word representation has shifted, reflecting stronger changes in language use for that particular lexical item. However, we can also probe the diachronic relation between a word embedding and the embeddings for the other words in the vocabulary, looking at whether relations changed coherently over time at the neighborhood level. For example, we can look at the similarity between *keyboard* and its most similar words in 1800 and the similarity between the same words today, measuring their overall difference. We can therefore probe different measures

of language change extracted from the same framework and investigate whether they relate differently with language acquisition.

While the theory behind our approach is well established, its feasibility crucially relies on recent improvements in temporal distributional models of language (Bianchi et al., 2020; Bowern, 2019; Di Carlo et al., 2019; Hamilton et al., 2016b; Tahmasebi et al., 2019). These methods have been used to investigate several phenomena in linguistics (including syntactic productivity, Perek, 2014; semantic, Sagi et al., 2011; Hamilton et al., 2016b; and morphosyntactic changes, Hilpert & Perek, 2015), history of culture (Hills, Proto, Sgroi, & Seresinhe, 2019; Li, Engelthaler, Siew, & Hills, 2019; Soni, Klein, & Eisenstein, 2021), language learning (Hills & Adelman, 2015), and computational social sciences (Garg, Schiebinger, Jurafsky, & Zou, 2018) just to name a few applications next to the direction we are pursuing in our work. Particularly relevant for our study, Hamilton et al. (2016a) quantified language change considering both item- and neighborhood-level measures and showed a dissociation between the two with respect to lexical categories, suggesting the two measures indeed capture different phenomena in language change.

To summarize, our work builds on previous studies highlighting how ontogenetic and ethnogenetic language dynamics are intertwined (Christiansen & Chater, 2008). We hypothesize that there exists a unique relation between the degree to which usage patterns change over time and the words that are learned earlier, such that words with stabler usage patterns over time tend to be learned earlier. In testing this hypothesis, we improve over previous studies by investigating the relation between language change and acquisition using a novel data-driven, corpus-based, scalable way of quantifying language change (Bianchi et al., 2020; Di Carlo et al., 2019), which is grounded in language use (Firth, 1957; Wittgenstein, 1953) and builds upon previous studies showing the effectiveness of similar approaches in addressing several linguistic phenomena (Hamilton et al., 2016a, 2016b; Hilpert & Perek, 2015; Perek, 2014; Sagi et al., 2011).

2. Methodology

2.1. Data

Several datasets were used to carry out this study. First and foremost, we used the Corpus of Historical American English (CoHA[4]; Davies, 2010) to track the evolution of the English language. This corpus contains more than 400 million words covering the period between 1810 and 2009 and is balanced by genre (fiction, magazine, newspaper, other nonfiction) for each decade. We split the corpus into six time slices to build time-dependent word embeddings.

Moreover, we retrieved *AoA* norms from the dataset constructed by Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012), which ensures a good coverage and has been widely used in the literature to track *AoA*. The validity of subjective estimates has been further confirmed by lab- and test-based results (Biemiller, Rosenstein, Sparks, Landauer, & Foltz, 2014; Brysbaert, 2017; Gilhooly & Gilhooly, 1980; Łuniewska et al., 2016; Morrison, Chappell, & Ellis, 1997), suggesting that these *AoA* norms are reliable. Other datasets were further used to derive

control variables, necessary to ensure that any relation we may detect involving language change and *AoA* is not best explained by known predictors of acquisition. We used the dataset released by Brysbaert, Warriner, and Kuperman (2014) to get concreteness norms and dominant part of speech (PoS) tag for a large sample of English words and extracted word frequency and contextual diversity[5] (Adelman et al., 2006) estimates from SUBTLEX-US (Brysbaert, New, & Keuleers, 2012), which has been shown to provide a very good fit to several psycholinguistics tasks (Baayen, Milin, & Ramscar, 2016). Orthographic neighborhood density was operationalized using Orthographic Levenshtein Distance (*OLD20*, Yarkoni, Balota, & Yap [2008]) and was computed using the OLD20 Python package[6] using the vocabulary from SUBTLEX-US as reference. In essence, *OLD20* averages the Levensthein distance between a target word and its 20 nearest orthographic neighbors.

2.2. Word embeddings

In order to quantify diachronic language change in usage patterns, we need to represent words in such a way that representations obtained over different slices are comparable. Standard word embeddings models such as*word2vec* (Mikolov et al., 2013) are not able to generate temporal word embeddings and to capture the temporal dimension of text: The stochasticity of neural networks prevents methods such as *word2vec* from being used to represent multiple slices of text (see Di Carlo et al., 2019, for further technical details).

Over the past few years, different methods have been proposed to align different temporal slices of the same corpus with the goal of quantifying lexical semantic shift in a data-driven and scalable way (Bianchi et al., 2020; Di Carlo et al., 2019; Dubossarsky et al., 2019; Hamilton et al., 2016b; Kim et al., 2014; Kulkarni et al., 2015; Rudolph & Blei, 2018; Tahmasebi et al., 2018; Yao et al., 2018). For example, Hamilton et al. (2016b) used Procrustes transformation to align embeddings, while Yao et al. (2018) used a joint optimization procedure. More recently, Di Carlo et al. (2019) proposed the temporally aligned word embeddings with a compass (TWEC[7]) model, which extends the continuous bag of words (CBoW) architecture (Mikolov et al., 2013). The CBoW architecture is a neural network with one hidden layer and uses two matrices to learn lexical representations, a *target* matrix and a *context* matrix. The TWEC model exploits this aspect of the CBoW architecture by first training a general embedding space using all the available text, ignoring the time dimension. This time-independent embedding space is the compass, that is, a general representation to which the other slices can be aligned. The context matrix of the compass is extracted and used to initialize (and freeze) the context matrix of a slice-specific CBoW model. This approach ensures that all slices share the same context matrix and that slice-specific embeddings are aligned. The TWEC model has been found to outperform competing models in aligning slices of text for temporal analogical reasoning (Di Carlo et al., 2019), meaning shift analysis (Bianchi et al., 2020), and narrative text understanding (Kanjirangat, Mellace, & Alessandro, 2020). Due to its simplicity, scalability and effectiveness in creating aligned word embeddings compared to previous methods (Hamilton et al., 2016b; Yao et al., 2018), the TWEC model is used in the current study to derive measures of change in temporal usage patterns of individual words.

2.3. Notation

In our current setting we have a collection T of sets of documents T_i, where $T = T_1 \cup T_2 \cdots T_t$, with t being the number of temporal slices considered. The vocabulary V consists of all the words in the corpus, while each slice is associated with a slice-specific vocabulary V_j; it follows that $V = \bigcup V_j$. The nth word of the V vocabulary in slice j is identified as w_n^j and the corresponding vector representation is identified in bold, $\mathbf{w}_n^j$.[8] However, note that in practice there is the possibility of a word not being present in a specific slice (e.g., the word *smartphone* is not present in texts from the 1800), and thus $\mathbf{w}_n^j$ might not be defined.

2.4. Quantifying language change

We use three different methods to quantify language change[9] (Gonen et al., 2020; Hamilton et al., 2016a). In order to ensure comparability across words, we only computed measures of language change for words that appeared at least 25 times in each of the six slices in which the corpus was divided, such that reliable lexical representations could be learned. As mentioned in Section 1, we used measures to quantify item-level as well as neighborhood-level language change. Item-level measures probe the change in the usage patterns of individual lexical items by comparing representations of the same word derived at different times. On the contrary, neighborhood-level measures consider the relation between a single lexical item and its similar items in the language, quantifying how this relation changed over time.

The first measure, vector coherence (*VC*), is an item-level measure that quantifies the coherence of the temporal word embeddings extracted from different time epochs and is implemented following Eq. 1. t is the number of slices in which the corpus is split and cos is the cosine similarity between two vectors. Importantly, we sum[10] the cosine similarity for pairs of embeddings not only for adjacent time slices but for all pairwise comparisons, to account for all possible trajectories of coherence and lack thereof. Higher values thus indicate a more coherent lexical representation over time:

$$VC(w_n) = \sum_{i,j \in T \text{ where } i \neq j} \cos(\mathbf{w}_n^i, \mathbf{w}_n^j). \tag{1}$$

Our second measure, local neighborhood coherence (*LNC*), addresses neighborhood-level change and tracks the coherence of the relation between a word embedding and the embeddings of its nearest semantic neighbors (Hamilton et al., 2016a). We describe its implementation in Eqs. 2 and 3. This measure is a second-order cosine similarity—given two corpus slices i and j, we compute the vector s_n^i as follows: We collect the k-nearest neighbors ($\mathcal{N}$) of both $\mathbf{w}_n^i$ and $\mathbf{w}_n^j$ and we then compute the similarity between $\mathbf{w}_n^i$ and all the neighbors in the embedding space i. Eq. 2 shows how to compute the values of the vector s^t. It can, however, be the case that a nearest neighbor of $\mathbf{w}_n^j$ does not occur in slice i, making it impossible to retrieve the corresponding embedding. When this happens, we compute the similarity between $\mathbf{w}^t$ and the average word embedding in slice t, since if no information is available from co-occurrences, the only available reference point is the average word embedding:

$$\mathbf{s}_n^i(x) = \cos(\mathbf{w}_n^i, \mathbf{w}_x^j) \; \forall w_x \in \mathcal{N}(w_n^i) \cup \mathcal{N}(w_n^j). \tag{2}$$

We repeat the same procedure for $\mathbf{w}_n^j$, obtaining two vectors $(\mathbf{s}_n^i, \mathbf{s}_n^j)$ of cosine similarities capturing how the same word relates to its nearest neighbors in two different time epochs. We then compute the cosine similarity between these two vectors of cosine similarities for all pairs of time slices as detailed in Eq. 3, summing pairwise similarities over different time slices to obtain our target measure *LNC*. Words keeping a coherent relation with their nearest semantic neighbors over time will have a higher *LNC*:

$$LNC(w_n) = \sum_{\forall i,j \in T \text{ where } i \neq j} \cos(\mathbf{s}_n^i, \mathbf{s}_n^j). \tag{3}$$

The third measure, *J*, is similar to *LNC* in that it tackles the neighborhood-level change but is fully symbolic, since it does not exploit the geometrical relations across word embeddings. Like *LNC*, this measure also tracks the coherence between local neighborhoods but it is computed through the Jaccard coefficient (hence the name) between the sets of nearest neighbors of $\mathbf{w}_n^i$ and $\mathbf{w}_n^j$, as detailed in Eq. 4. The Jaccard coefficient is a measure of set overlap obtained by dividing the cardinality of the set intersection by the cardinality of the set union. A coefficient of 1 indicates perfect overlap, while a coefficient of 0 indicates no overlap. In the case at hand, the Jaccard coefficient between two words reflects the normalized overlap between the set of their neighboring words in two corpus slices. As for *VC* and *LNC*, we compute *J* for all possible pairwise comparisons of time slices and sum the pairwise scores. Words with a high *J* will have more coherent local neighborhoods over time, disregarding the geometrical relation between the target word and its neighbors:

$$J(w_n) = \sum_{\forall i,j \in T \text{ where } i \neq j} \frac{\mathcal{N}(w_n^i) \cap \mathcal{N}(w_n^j)}{\mathcal{N}(w_n^i) \cup \mathcal{N}(w_n^j)}. \tag{4}$$

The main difference between *J* and *LNC* is that the former only considers the identity of the neighbors and tracks how many occur both at time i and at time j over how many unique neighbors occur at both time epochs. Therefore, the problem of missing embeddings is avoided. For the purpose of computing both *LNC* and *J*, we set $k = 25$ as the number of nearest neighbors being retrieved, in line with Hamilton et al. (2016a).

Table 1 summarizes the different measures of language change considered in this work, listing the abbreviations, which will be used throughout the paper, their names, and a short description.

For every measure of language change, we also compute a quasi-random control measure to ensure that any correlation between language change and *AoA* does not occur by chance or due to properties of the embedding spaces that are irrelevant to capturing language change. The random counterpart of *VC*, *rVC*, is computed by randomly sampling a word embedding among the 10 nearest neighbors of $\mathbf{w}^i$ and $\mathbf{w}^i$ itself, repeating this procedure for $\mathbf{w}^j$, and computing the cosine similarity between the two randomly sampled embeddings. Therefore, we expect *rVC* to have a weaker but qualitatively similar relation with *AoA* to *VC*. *rJ* was computed similarly, randomly sampling a word embedding among the 10 nearest neighbors of $\mathbf{w}^i$ and $\mathbf{w}^i$ itself, getting its k nearest neighbors, repeating this procedure for $\mathbf{w}^j$, and computing

Table 1
Measures of language change used in this work

Abbr.	Measure	Description
VC	Vector coherence	Tracks the consistency of each lexical representation by measuring the cosine similarity between the embeddings of the same word in different corpus slices
LNC	Local neighborhood coherence	Tracks the consistency in the cosine similarity between a word embedding and its k nearest neighbors in two different corpus slices
J	Jaccard coherence	Tracks the consistency in the identity of the nearest neighbors of a target word using the Jaccard coefficient as a measure of normalized set overlap in two different corpus slices

Note. The table lists the abbreviation (Abbr.) in the first column, the full name (Measure) in the second, and a short description in the third.

the Jaccard coefficient between the set of neighbors at time *i* and the set of neighbors at time *j*. *rLNC*, on the contrary, was computed by sampling words at random from the whole vocabulary, which is less problematic since what is being tracked is the relation between the target word and the other words in the language. It is therefore conceivable that this relation may be coherent even when computed over words that fall outside the local semantic neighborhood of the target word.

Finally, it is known that word embeddings also capture corpus frequencies (Schakel & Wilson, 2015). Therefore, we computed a measure of diachronic frequency change in the CoHA to be able to tease apart the relation that changes in usage patterns may have with *AoA* from that of mere frequency variations. For every temporal slice, we computed the frequency per million (*FpM*) of each token and summarized frequency variations summing *FpM* values over corpus slices, in line with what we did for the target measures of language change. Words with high *FpM* thus tend to be more frequent across the whole corpus.

2.5. Statistical approach

All analyses were performed in the R programming language (R Core Team, 2017). All independent variables were first transformed using a Box–Cox power transformation (performed using the package *MASS*, Venables & Ripley, 2002) to remove the skew in their distribution and then *z*-transformed. This procedure ensures that all variables exist on the same scale, that regression coefficients can be directly compared, and that their distribution is as close as possible to normal across all analyses. We first fitted a baseline statistical model, where *AoA* is modeled as a linear combination of frequency, contextual diversity (Adelman et al., 2006), concreteness, *OLD20* (Yarkoni et al., 2008), length in characters, dominant PoS tag, and diachronic frequency. We then added each measure of semantic change to this baseline model and assessed whether its fit improved by considering the difference in the Akaike Information Criterion (ΔAIC) between the baseline and each model including a measure of language change. The choice to use *AoA* as our dependent variable is primarily motivated by the necessity to control for variance explained by diachronic shifts in frequency,

to ensure that the potential relation between changes in usage patterns and *AoA* can be really ascribed to an effect of stability in usage patterns and is not reducible to an epiphenomenon of other diachronic patterns in the corpus. In the discussion, we consider the possible theoretical implications of this approach.

3. Results

We report and discuss results obtained with word embeddings with 40 dimensions and training the TWEC model with a window size of three words to the left and right of the target word. Crucially, we investigated other values for these hyperparameters, finding consistent patterns these analyses are discussed at length in the Appendix. We preprocessed the text using Spacy,[11] removed stopwords and punctuation, and lemmatized the corpus. Finally, we removed all the words that appeared 10 times or fewer in a slice to reduce the noise in the representations. This entails that a word may be preserved in a slice where it occurs 20 times but discarded in a slice where it only occurs 5 times.

3.1. Correlation analysis

We start by considering the quantitative relations concerning language change, *AoA*, and the set of chosen control variables. The correlation matrix is shown in Fig. 1. First, we see strong positive correlations across different measures of language change, suggesting that they largely capture similar patterns.

Fig. 1 also suggests that there is a relation between *AoA* and measures of language change, particularly looking at *VC* and *J*. We see that *J* has the strongest correlation with *AoA* norms ($r = -0.334$), followed by *VC* ($r = -0.306$), while *LNC* has the weakest correlation ($r = -0.182$). Therefore, both item- and neighborhood-level measures of language change seem to correlate with *AoA*, although to a lesser extent than other predictors. The analysis reported in the following section investigates whether there is unique variance in *AoA*, which is captured by measures of language change. This analysis becomes particularly relevant when looking at the collinearity between measures of language change and other covariates; in particular, the correlation between *VC* and *J* on the one hand, and concreteness and frequency measures, both synchronic and diachronic, on the other may signal that the effect of *VC* and *J* could be reduced to a frequency or concreteness effect. On the contrary, *LNC* seems to be largely orthogonal to other variables, even though its contribution in explaining *AoA* is not very large to begin with.

Turning to quasi-random control measures, *rVC*, *rJ*, and *rLNC*, we observe smaller correlations between *AoA* and all three, confirming that true measures of language change have stronger correlations with *AoA* than the quasi-random counterparts, although the difference is small for *LNC* and *rLNC*. Moreover, whereas *VC* and *J* have stronger correlations with frequency values than their quasi-random counterparts, the opposite is observed for *LNC*. Quasi-random measures, however, tend to be more correlated with concreteness.

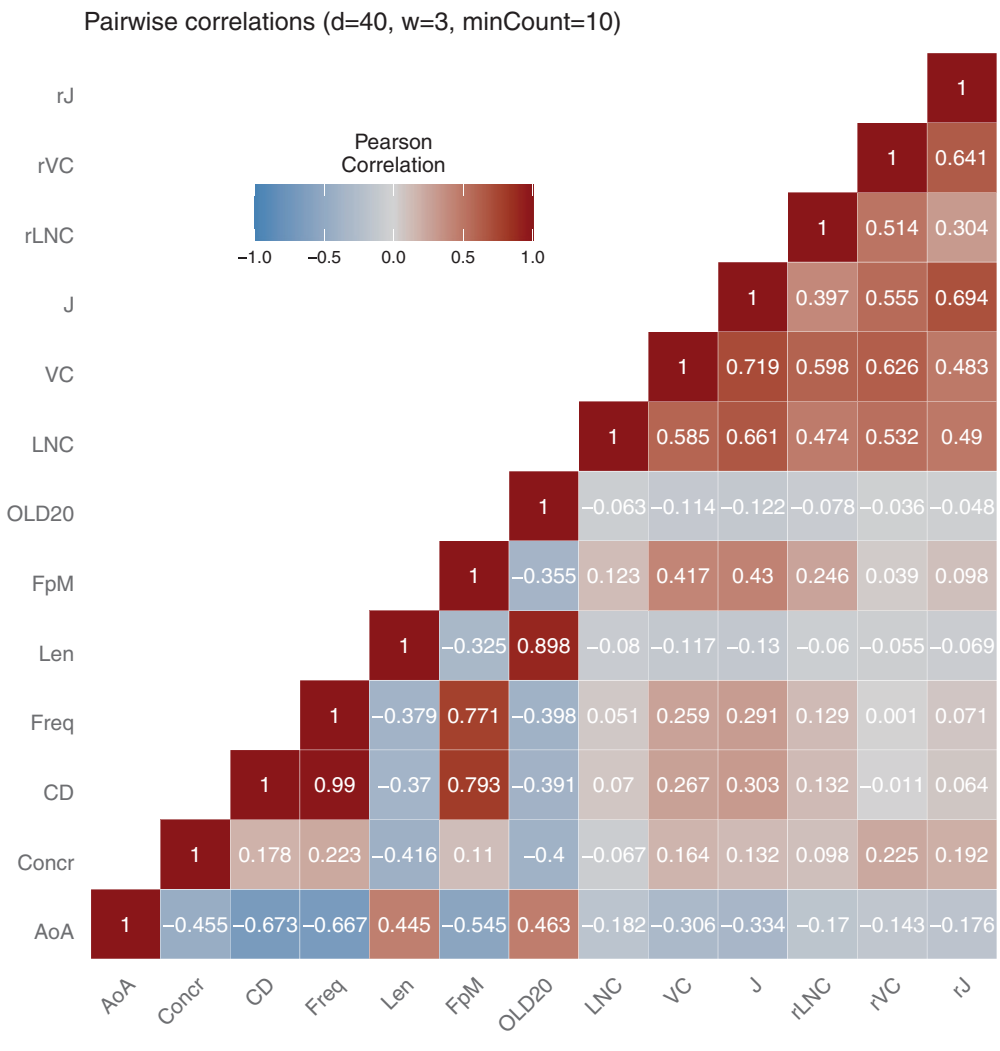

Fig. 1. Pairwise correlations between measures of language change, *AoA*, and control variables (TWEC with dimensionality = 40, window size = 3, and enforcing a minimum count = 10 on word frequency in a slice). Abbreviations: *LNC*, local neighborhood consistency; *VC*, vector coherence of word embeddings; *J*, Jaccard coefficient of word semantic neighbors; *rLNC*, random local neighborhood consistency; *rVC*, random vector coherence of word embeddings; *rJ*, random Jaccard coefficient of word semantic neighbors; *AoA*, age of acquisition; *Concr*, concreteness; *Freq*, frequency (SUBTLEX); *CD*, contextual diversity (SUBTLEX); *Len*, length in letters; *FpM*, frequency per million (CoHA).

3.2. Multiple regression analysis

As mentioned in Section 2.5, we first fitted a baseline multiple linear regression model predicting *AoA* using word frequency, contextual diversity, orthographic word length, concreteness, dominant PoS tag, *OLD20*, and *FpM* to control for diachronic frequency variations

Table 2
Summary statistics of $lm(AoA \sim basemodel + (Predictor))$, presenting the β coefficient, standard error (SE), t statistic, p value, r^2, AIC, and ΔAIC with respect to the baseline statistical model provided at the top

Predictor	β	SE	t	p	r^2	AIC	ΔAIC
Base					0.599	32,414.10	0
LNC	−0.3844	0.0188	−20.500	<.001	0.618	32,005.17	408.93
rLNC	−0.1425	0.0195	−7.309	<.001	0.601	32,362.72	51.38
J	−0.3059	0.0208	−14.677	<.001	0.609	32,202.91	211.19
rJ	−0.2090	0.0192	−10.873	<.001	0.604	32,298.44	115.66
VC	−0.2536	0.0290	−12.093	<.001	0.606	32,270.79	143.31
rVC	−0.2203	0.0194	−11.384	<.001	0.605	32,287.20	126.90

in the source corpus.[12] Then, we separately added each measure of language change to this baseline statistical model to assess their unique contribution. Results are reported in Table 2. The measures of fit for the baseline statistical model are provided for reference at the top.

We see that all models including measures of language change have a lower *AIC* and a higher r^2 than the baseline statistical model. Moreover, we see that measures of language change have lower *AIC* scores than their corresponding quasi-random counterparts, confirming that true measures of language change explain more unique variance in *AoA* norms. In detail, we see that the highest improvement in fit is brought by *LNC* ($\Delta AIC = -408.93$), with *J* ($\Delta AIC = -211.19$) following and *VC* ($\Delta AIC = -143.31$) bringing the least improvement. Considering the collinearity between *VC* and the control variables, especially concreteness and frequency measures, it is not surprising that its effect becomes weaker once control variables are included in the statistical model. Similarly, it is not surprising to see that *LNC* explains the largest share of unique variance in *AoA* considering that it is largely orthogonal to other predictors. Finally, looking at β coefficients, we see that for a 1 standard deviation increase in *VC*, the predicted *AoA* drops by a quarter of a year ($\beta = -0.2536 \pm 0.0290, t = -12.093$). A 1 standard deviation increase in *LNC* results in a decrease in predicted *AoA* of about 5 months ($\beta = -0.3844 \pm 0.0188, t = -20.500$). Finally, a 1 standard deviation increase in *J* results in a drop of predicted *AoA* of about a third of a year ($\beta = -0.3059 \pm 0.0208, t = -14.677$).

The regression analysis summarized in Table 2 confirms that the pattern of the relations between different measures of language change and *AoA* does not hold once standard covariates of *AoA* and diachronic frequency variations are controlled for. Although *J* and *VC* had a stronger correlation with *AoA* than *LNC*, we see that this pattern is reversed in the regression analysis. This suggests that a large portion of the variance that *VC* and *J* explain in *AoA* is actually better accounted for by other covariates. To verify this, we regressed *AoA* on *VC* alone, reporting a coefficient almost three times as large as that reported in Table 2 ($\beta = -0.8215 \pm 0.0281, t = -29.23$). We repeated the same procedure for *J* and obtained a coefficient which was almost three times as large as that provided in Table 2 ($\beta = -0.8975 \pm 0.0278, t = -32.26$). On the contrary, the coefficients of the control variables in the baseline model remain largely unaltered by the inclusion of *VC* and *J*. This pattern

is likely explained by the pairwise correlations between *VC* and *J,* on the one hand, and frequency, contextual diversity, and concreteness, on the other.

3.3. Examples of words with high and low diachronic coherence

After having shown that there indeed is a unique relation between different measures of language change and acquisition patterns, it is interesting to look at which words have the highest and lowest diachronic coherence as measured by *VC*, *LNC*, and *J*. The examples we discuss show that, while they largely agree, different measures of language change capture different subtle patterns. This suggests that language change is a composite phenomenon that cannot be easily reduced to a single measure without missing some of its relevant aspects.

Looking at the words with a high *VC*, we see several words referring to body parts (*finger*, *hand*, *wrist*, *arm*, and *hair*). Moreover, we observe basic-level concepts, referring to entities, for example, *water*, *word*, *wind*, *hill*, *dress*, *sea*, *bird*, and *church*, and actions, for example, *sing*, *kill*, and *buy*. Time expressions, for example, *year*, *evening*, *month*, *winter*, *summer*, and *morning*, also tend to show high *VC*. Finally, we also observe words with very specialized meanings having high *VC*, for example, *embroider*, *guttural*, *foliage*, *moisture*, *complexion*, *symptom*, and *dialect*, suggesting that words that are used in specific domains with strongly conventionalized meanings tend to have high diachronic coherence. These words, however, are also of low frequency; their stability is in line with and further qualifies the evidence provided by Monaghan and Roberts (2019), who showed that low-frequency words have a rather low probability of being replaced. Our analysis suggests that domain specificity may be a further factor influencing diachronic stability.

On the other hand, many words with low *VC* appear to have been at the center of technological changes. Examples are *projector*, which came to refer to a tool used to show videos while it was used to refer to a person who made plans; *monitor*, which used to refer to people overseeing tasks while now primarily indicates video terminals; *stainless*, which shifted from a moral connotation to referring to steel; *console*, which came to indicate a device to play video games next to the action of comforting somebody; a highly polysemous word such as *recorder* that could indicate a wind musical instrument, a judge, and now primarily a tool for audio recording. We also observe the effect of cultural changes on lexical meanings in words relating to the sexual sphere such as *hooker*, *abortion*, *erection*, and *pregnant*. Finally, we see many words now relating to gay culture whose meaning changed over time, often words repurposed to stigmatize or insult gay people. While the very word *gay* does not have one of the lowest *VC* values, words such as *aids*, which came to indicate the disease as a lexicalized acronym, *fag*, which used to refer to cigarettes, or *faggot*, which indicated sticks of wood tied together, all appear at the bottom of the list when considering *VC*.

Fig. 2 graphically shows the *VC* of four words, two with high coherence (*finger* and *thunder*) and two with low coherence (*pregnant* and *recorder*), which highlight two different trajectories of diachronic change. On the one hand, *pregnant* shows two distinct clusters, signaling a shift in usage between 1880 and 1920; on the other hand, *recorder* shows a more constant shift. On the contrary, the temporal vector representations for *finger* and *thunder* are much closer in space to each other, signaling an overall stability in language use.

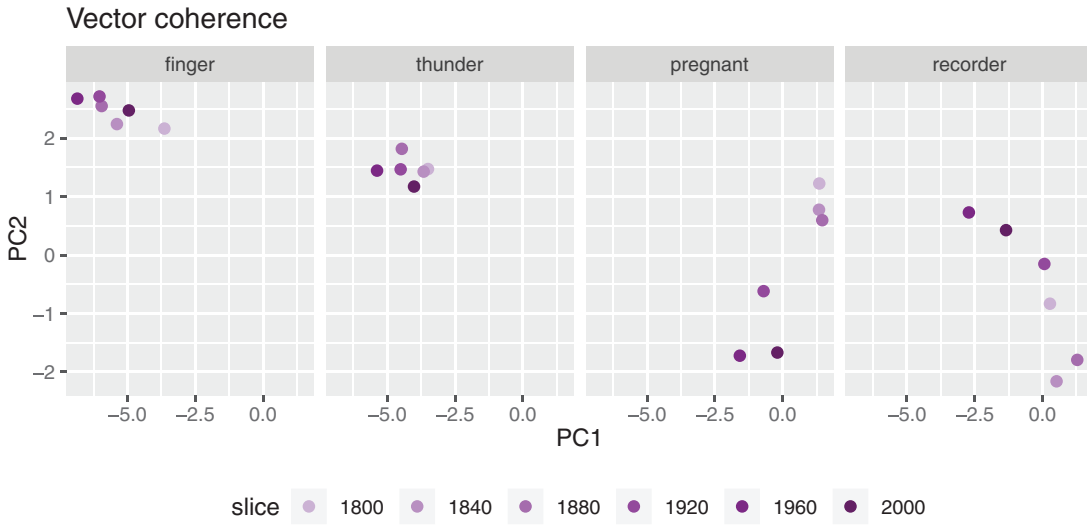

Fig. 2. Vector coherence . Each panel represents a different word (from left to right: *finger*, *thunder*, *pregnant*, *recorder*. Each dot represents the coordinates of a temporal vector representation extracted from a specific slice of the CoHA and mapped onto a two-dimensional space obtained by applying Principal Component Analysis to the original 40-dimensional vectors (the first principal component [PC1] is the *x*-axis, while the second principal component [PC2] is the *y*-axis). Each color indicates a different temporal slice.

Moving to measures of neighborhood-level coherence, all words mentioned as having low *VC* also appear to have low *LNC* and *J* values, suggesting that words that underwent drastic changes in language use did so at both item and neighborhood levels. It is therefore interesting to check whether the same happens for words with high *VC*, that is, whether words with stable vector representations also top the chart in terms of measures of neighborhood-level coherence.

Among words with high *LNC*, we no longer observe basic-level concepts. We still see some words referring to body parts, for example, *face*, *tooth*, *eyelid*, and *eye*, and temporal expressions, for example, *night* and *twilight*. Both patterns, however, are weaker than observed for *VC*, and words appear to be less prototypical. Words with high *LNC* include animals, such as *snake*, *cat*, *vulture*, *raccoon*, *toad*, *dog*, and *frog*, and fruit, such as *melon*, *pumpkin*, *cabbage*, and *turnip*. Two very strong patterns among words with high *LNC*, which sometimes overlap, involve words including nonarbitrary relations between form and meaning, on the one hand, and words relating to the broad semantic field of suffering and sadness, on the other. The first trend is exemplified by words including phonaestemes, such as *fling*, *sneer*, *glimmer*, *glitter*, and *snore*, and by phonosymbolic words, such as *shriek*, *howl*, *wail*, *roar*, *moan*, *whine*, *groan*, *rumble*, *creaking*, *sob*, and *squeak*. The second trend is represented by some of the previously mentioned words (*shriek*, *howl*, *moan*, *groan*) and others such as *grief*, *remorse*, *suffering*, *anguish*, *humiliation*, *agony*, or *sorrow*.

Fig. 3 shows a word with high *LNC*, *evening*, and a word with low *LNC*, *aids*. The plots show the cosine similarity between the target word and the word's nearest neighbors in two

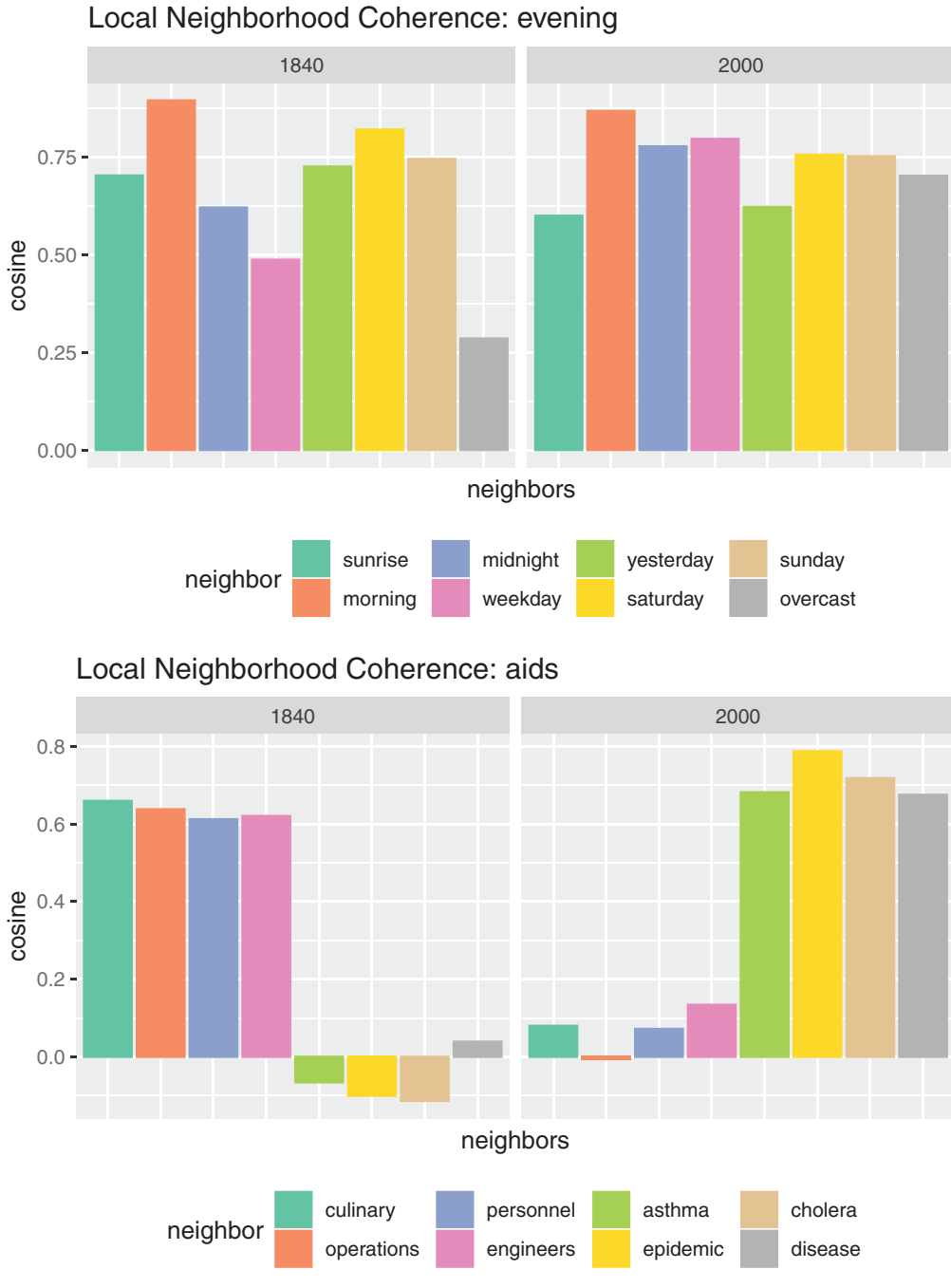

Fig. 3. Local neighborhood coherence. The top panel represents the similarity patterns of the word *evening* while the bottom panel represents the similarity patterns of the word *aids*. The *x*-axis shows eight of the nearest neighbors taken from two different temporal slices (i.e., for each word we take four nearest neighbors per slice), the *y*-axis indicates the cosine similarity between each neighbor and the target word, while the facets indicate two temporal slices, 1840 on the left and 2000 on the right.

Table 3
Examples of words with high and low *J*

Target	Neighbors (1840)	Neighbors (2000)	*J*
Fifth	**sixth**, **fourth**, **eighth**, **ninth**, **tenth**, **seventh**, twentieth, **twelfth**, thirtieth, fourteenth, **second**, fortieth, inclusive, fiftieth, **eleventh**, thirteenth, **fifteenth**, sixteenth, thirty, seven	**fourth**, **seventh**, **sixth**, **ninth**, **eighth**, **tenth**, 12th, 13th, **eleventh**, 11th, 14th, 33rd, 10th, **fifteenth**, consecutive, **twelfth**, 34th, 15th, 16th, **second**	5.39
Aids	engineering, auxiliary, superintendence, tactics, desideratum, adjunct, culinary, housewifery, procurement, ception, ose, operations, commissariat, offi, reconnaissance, portant, ciety, engineers, cult, department	HIV, std, epidemic, trachoma, syphilis, malaria, Hodgkin, cholera, communicable, polio, cancer, Alzheimer, lupu, Ebola, lupus, asthma, ovarian, testicular, tuberculosis, colon	0.22

Note. The first column (*Target*) contains the target word, the second column (*Neighbors (1840)*) shows the 20 nearest neighbors of the target word in the time slice ending in 1840, the third column (*Neighbors (2000)*) shows the 20 nearest neighbors of the target word in the time slice ending in 2000, and the last column provides the overall *J* score. The shared neighbors in the two slices are in bold. Words like *ception*, *ose, offi, and cietyare likely errors of the Optical Character Recognition system used to digitise the corpus or errors of the lemmatiser*.

different time slices, 1840 and 2000. Of all the neighbors, we picked eight to avoid cluttering the plot and ensure to make the patterns visible. Looking at *evening*, we see that all neighbors have kept a rather consistent similarity with the target word, except for *overcast* and *weekday*, which became more similar to the target in 2000. The fact that only minor differences appear in the bar heights signals that the relation between *evening* and its nearest neighbors in different corpus slices remained largely consistent diachronically, contributing to an overall high *LNC*. On the contrary, in the bottom panel, we observe that the relation between the target word *aids* and some of its neighbors changed drastically. Although words with similar usage patterns in 1840 included *culinary* and *personnel*, in 2000 the word was being used more similarly to *epidemic* or *disease*. The plot shows how the bars change, with neighbors with a high similarity in 1840 falling near 0 in 2000 and vice versa.

Moving to the third measure of language change, among words with high *J*, we again observe time expressions such as *night*, *hour*, *day*, *morning*, *ago*, *evening*, and *afternoon*, numbers and ordinals, such as *seven*, *tenth*, *fourth*, *sixth*, *fifth*, *twice*, *eighth*, and *ninth*, as well as words referring to kinship relations, such as *child*, *daughter*, and *wife*. These trends suggest that *J* is highest for words from the same closed, narrow semantic domains, where words are likeliest to be neighbors of each other. This highlights a potential limitation of a purely symbolic measure, which could equate diachronic coherence to closed semantic domains. Moreover, we observe other trends which were also reported when discussing words with high *VC* and high *LNC*, including words referring to suffering (*sob*, *grief*), words including nonarbitrary form-meaning relations (*scream*, *howl*, *moan*, *roar*, *shriek*), body parts (*forehead*, *hand*, *finger*), fruit (*cabbage*, *melon*), and basic-level concepts (*window*, *table*, *walk*, *door*).

Table 3 shows one word with high *J* (*fifth*) and one with low *J* (*aids*). Again, for illustration purposes we only show two time slices, 1840 and 2000. We see how *aids* does not share any

neighbor in distant time slices, while *fifth* shares several neighbors even after several decades. It is interesting to see, however, that this measure is very strict in that it only accepts as valid overlap the occurrence of the exact same neighbor. When looking at the neighbors of *fifth,* we see that the overlap could have been much higher if the convention of writing ordinals in letters rather than with numbers would have stuck: In 1840 we see the token *fourteenth,* while in 2000 we see the token *14th*. However, venturing into processes of normalization would have introduced further degrees of freedom in the modeling framework which fell outside the scope of the current paper.

4. Discussion

In this study, we report evidence of a unique relation between language change and *AoA*, which remains after controlling for variables, which are known to influence the age at which words tend to be acquired as well as usage-based measures of diachronic variation. All else being equal, words with more coherent diachronic usage patterns tend to be acquired earlier. Our results were obtained leveraging temporal word embeddings derived using the TWEC model (Di Carlo et al., 2019) and with the CoHA (Davies, 2010), a large diachronic corpus of American English spanning two centuries.

We improved over previous studies investigating the relation between language evolution and acquisition dynamics (Monaghan, 2014; Monaghan & Roberts, 2019; Vejdemo & Hörberg, 2016) by targeting a larger and more representative sample of words. We achieved this by using an entirely data-driven, unsupervised corpus-based approach to quantifying language change, which is grounded in diachronic language use (Firth, 1957; Harris, 1954; Wittgenstein, 1953) rather than exploiting measures of language change, which require expert annotations or carefully constructed resources (Monaghan, 2014; Monaghan and Roberts, 2019; Pagel et al., 2007; Vejdemo and Hörberg, 2016; Winter et al., 2014). This approach builds on recent advancements in the study of language change using distributed temporal word embeddings (Bianchi et al., 2020; Di Carlo et al., 2019; Hilpert & Perek, 2015; Kim et al., 2014; Kulkarni et al., 2015; Perek, 2014; Sagi et al., 2011; Yao et al., 2018). Finally, we implemented and evaluated three different ways of capturing language change, which target both item- and neighborhood-level patterns of change in usage patterns (Gonen et al., 2020; Hamilton et al., 2016a), offering a more thorough and insightful characterization of the phenomenon at hand.

We also extended the extant literature on temporal word embeddings (Kutuzov et al., 2018; Tahmasebi et al., 2018) by applying the TWEC model, a novel method to align diachronic lexical representations (Di Carlo et al., 2019), to the analysis of acquisition patterns and providing further evidence that measures of language change developed in the NLP community can be fruitfully used to investigate cognitive phenomena. Moreover, we showed that different measures of language change not only relate differently with lexical categories (Hamilton et al., 2016a) but also with word-level properties and language acquisition. *VC*, for example, was reported to correlate more with frequency and concreteness, such that more concrete and frequent words tend to have more coherent embeddings over time. This result fits with

evidence reported by Vejdemo and Hörberg (2016) that more imageable words tend to undergo less lexical replacement over time, thus being more stable from a lexical point of view. A distributed measure of neighborhood coherence such as *LNC*, on the contrary, is less correlated with frequency and concreteness, and explains more unique variance in acquisition patterns.

In detail, the correlation analysis showed that all measures of language change correlated negatively with *AoA*, indicating that words with stabler diachronic usage patterns tend to be acquired earlier. The regression analysis characterized relations further. The effect of *VC* on *AoA* reduced the most when controlling for other variables. Measures of neighborhood coherence, *LNC* and *J*, on the contrary, explained more unique variance after controlling for covariates of language change and other word-level properties with a known influence on *AoA* (Braginsky et al., 2019; Hills et al., 2010). This dissociation between measures of item- and neighborhood-level coherence suggests that, once other properties of target words are taken into account, the relation between *AoA* and language change emerges more clearly at the neighborhood level, where the relations between different lexical items over time are considered.

The qualitative analysis also highlighted interesting trends among words with higher vector or neighborhood coherence. In particular, we observed that words with high *VC* tend to refer to basic-level concepts, with higher frequency and concreteness (as confirmed by the positive correlations between *VC* and concreteness reported in Fig. 1). Words with both high vector and neighborhood coherence, on the contrary, tend to refer to time-related concepts, body parts, and numerals. The fact that we observe peculiar effects of language change (or better lack thereof) for number expressions in word embeddings is further evidence that language encodes important aspects of numbers and can inform the way we represent them cognitively (Rinaldi & Marelli, 2019). Our results show that number words tend to maintain coherent word representations over time, suggesting a central role in language, which should be further studied considering synchronic and diachronic patterns.

These pieces of evidence indicate, indeed, a unique and robust relation between stability in usage patterns and *AoA*, which holds after controlling for several known word-level predictors of acquisition and usage-based covariates of changes in usage patterns. With this conclusion, we complement previous studies that uncovered a relation between acquisition and language change (measured as a word's stability of form and its probability of being borrowed), while controlling for the effect of known predictors of language change (Monaghan, 2014; Monaghan & Roberts, 2019). Ours remains, however, a correlational analysis, which prevents us from drawing definitive conclusions about the causal relation between either dimension. How to describe this relation, then? A direct link between the two variables is unlikely. On the one hand, children are not exposed to the history of the language and so diachronic stability cannot directly affect *AoA*. On the other hand, *AoA* patterns estimated from the intuitions of adults in the 2010s cannot have exerted an influence on patterns of change which happened over the previous centuries, unless we assume that *AoA* estimates obtained today also capture *AoA* patterns in the past. The relation is thus likely to be more complex than a directional causal link. We get back to this point later in the discussion.

Even if we cannot directly interpret our results as arising from a direct relation between language change and *AoA*, previous studies can inform a more insightful discussion of the unique relation we uncovered. First, Newberry, Ahern, Clark, and Plotkin (2017) provided evidence of a negative correlation between frequency and language drift, such that more frequent words are less subject to stochastic drift in language transmission. Considering the robust negative relation between frequency and acquisition (Braginsky et al., 2019) as well as the negative correlation between our measure of change in frequency over time and *AoA*, a positive relation between drift and acquisition may be hypothesized, such that early acquired words are also less subject to stochastic drift and stabler over time. On a related but different level, it can be questioned whether children are at all sensitive to small changes that happen on a much longer time scale than that in which they learn a language. This concern can be addressed by considering evidence provided by Thompson, Kirby, and Smith (2016), who have shown how cultural transmission amplifies weak linguistic biases over time, testifying to the possibility that small changes in language use over a long period of time can build up and influence the language that is learned at a given time, and, we hypothesize, how this process happens.

However, by explicitly controlling for diachronic frequency variations in our regression analysis, our results also show that there is something unique about the relation between stability in usage patterns and *AoA*. We suggest that the relation at the heart of our study is best understood by considering the features of the language network, also considering that the strongest relation between change and acquisition emerged when quantifying change at the neighborhood level. Previous research has shown how this network consists of a few hubs with many connections and a lot of nodes with few connections (Steyvers & Tenenbaum, 2005). Such a network could form following different dynamics. One possibility is that new concepts are added to the network as a function of the number of connections each node in the network already has, with a rich-gets-richer effect (Steyvers & Tenenbaum, 2005). Under this hypothesis, words with more stable usage patterns could end up having more connections with other nodes in the network and act as hubs around which the network grows. However, this account has been challenged as a model of how individual children learn words (Hills et al., 2009, 2010). An alternative view posits that early learned words are more contextually diverse. Relating our finding to contextual diversity is harder since we explicitly control for it in our model. However, our analysis did show a moderate positive correlation between contextual diversity and *VC* and *J*, suggesting that words with more stable usage patterns tend to co-occur with a higher number of unique contexts. This fits with the hypothesis that words with a higher contextual diversity may play a role in structuring the growth and end state of language networks. That said, it is important to point out that the notion of contextual diversity has recently been called into question by Hollis (2020), who showed how it could boil down to a transformation of frequency. Measures that better disentangle the effects of frequency and contextual diversity will be needed in order to improve our understanding of these dynamics.

A third account of how language networks could grow during learning involves the relation of new nodes with already known words (Hills et al., 2009). New words do not enter the network earlier because they are attracted by hubs, but rather because unknown words are themselves hubs, being associated with several words in the environment (Hills et al., 2010).

The diachronic stability of lexical representations could play a crucial role in maintaining such dense relations in the environment and ultimately facilitating learning by structuring the lexicon in a convenient way. While our results cannot adjudicate between competing accounts of how individual language networks grow, they highlight how the structure of such networks is not simply constrained by the learning environment in which the child grows, but may depend on the long process that shaped the language network that children will eventually learn. Baumann (2018) highlights how changes that optimize learnability are less likely to get lost due to random fluctuations, which is instead a likelier outcome for innovations that optimize usability. This fits again with the evidence about stochastic drift offered by Newberry et al. (2017). The observation by Hills et al. (2009) that "the structure of information influences early [word] learning" fits nicely with our results, which highlight how the long process that molds the language network could make it easier to learn. This perspective also suggests that there likely is not a single directional causal relation between acquisition and change, with the current state of the language network being influenced by diachronic patterns, influencing learning, and in turn affecting diachronic patterns to come. The fact that a unique relation involving acquisition and change has been documented regardless of the directionality of the hypothesis, and thus after controlling for variables known to affect change and variables known to affect acquisition is a further hint of a feedback-loop kind of relation between these two constructs, where stability in usage pattern and early *AoA* go hand in hand and influence each other in cultural transmission of language. Early acquired words gain representation strength that guards against change (Monaghan, 2014; Monaghan & Roberts, 2019), and stability in usage patterns contributes to structuring the language network in such a way that some words are easier to learn than others. Further studies are needed, however, to probe the relation between psycholinguistic properties of language and its subjective experience (captured by *AoA* estimates in our study), on the one hand, and the properties of how the language system changes over time, on the other, to uncover whether this relation is indeed best characterized by a directional influence of one over the other or by mutual influences at different time scales.

At a more speculative level, diachronic stability may also relate to synchronic stability, such that words with more stable diachronic usage patterns are also used consistently during development. This could reinforce the information structure in the world and facilitate word learning (Brysbaert & Ghyselinck, 2006; Ghyselinck et al., 2004; Hills et al., 2009). The neural-network model of language learning proposed by Ellis and Lambon Ralph (2000) is particularly interesting in this respect. It predicts that early acquired words will exert a larger influence on the whole system because at the initial stages the model is still changing rapidly while it tends to settle later on, such that the connections in the network are primarily shaped when the network has higher plasticity. If early acquired words had unstable usage patterns, it would be harder to learn reliable connections, since new encounters with a word in context would push the learned representation in different directions. This is, however, a hypothesis that needs, at present, to be further tested. Recent models that learn context-dependent word embeddings (Devlin, Chang, Lee, & Toutanova, 2019; Peters et al., 2018) could be useful to quantify cross-sectional stability of representation and verify whether this relates to *AoA* on top of diachronic stability, or whether one explains all the variance the other does and some.

A further possible explanation of the reported relation between language change and acquisition patterns involves a third variable, which may affect both *AoA* and language change, that is, conceptual stability. This hypothesis rests on the observation of which words show high *VC*, *J*, and *LNC* values, on the pairwise correlation between *VC* and concreteness, and on the observation that the relation between *VC* and *AoA* is greatly reduced after controlling for concreteness. Therefore, it may be the case that words with more stable usage patterns refer to more stable concepts, which remain stable because they capture important and invariant aspects of reality (that are less subject to cultural evolution) and thus are also learned earlier. It is not unlikely that concepts, both concrete (such as body parts and fruit) and abstract (time expressions and numbers), whose referents are diachronically more stable, will be referred to more coherently in language and acquired earlier. In this sense, our purely linguistic measure of representational stability would be an epiphenomenon of a broader-level conceptual stability. There are, however, several models that build conceptual representations combining linguistic and modality-specific information, such as images (Bruni, Tran, & Baroni, 2014; Kiela & Bottou, 2014) and natural sounds (Kiela & Clark, 2015; Lopopolo & van Miltenburg, 2015). If datasets of images and sounds pertaining to a same concept from different time periods were available, the language-based model we considered could be extended to overcome this limitation and disentangle the contribution that pure stability of referent and pure stability in language use has on *AoA*. Moreover, even though it is debatable whether language use reflects meaning or meaning reflects use (Wittgenstein, 1953) and is thus hard to pin down the causal links between conceptual and linguistic stability, our work shows that there is a tight relation between concepts and the language, which is used to refer to them over time, which can be captured by temporal word embeddings, and can be used to investigate acquisition dynamics.

One more aspect of how word embeddings are learned in the CBoW architecture, and hence by the TWEC model, is worth discussing, particularly in the light of evidence provided by Vejdemo and Hörberg (2016). This study showed that polysemy as measured by number of WordNet senses had a negative relation with the rate of lexical replacement, such that more senses resulted in a lower probability of lexical replacement (Vejdemo & Hörberg, 2016). As we discussed in Section 1, distributional representations are derived by tracking co-occurrences. If a word is polysemous, the resulting word embedding will collapse all the different usages capturing the more frequent and distinctive ones better, such that the embedding for *keyboard* will reflect its shift in language usage from the music to the technology domain but still be shaped by all of its senses, technology- and music-related. If a word is entirely repurposed or is attached to a new meaning, it becomes more entrenched in the lexicon. Then it makes sense that its probability of being replaced decreases, as observed by Vejdemo and Hörberg (2016), since, after being repurposed or extended to new meanings, replacing it entirely would cause larger disruptions in the lexicon. At the same time, results provided by Winter et al. (2014) show that more polysemous words are likelier to be the origin of language change, thus of being repurposed or extended to a new meaning, making them more entrenched in the lexicon, and at the same time more likely to determine a higher rate of change as measured by our model. The observation that polysemy has a different influence on different measures of language change, correlating with lower rates of

replacement and stronger changes in usage patterns, calls for future studies to analyze how usage patterns relate to the rate at which word forms are introduced in or disappear from a language, and investigate a possible mediation of *AoA* in the relation between stability of usage and stability of form. If *AoA* were not just a proxy for representational stability (Juhasz, 2005; Monaghan, 2014), it would be expected to have a unique effect on stability of form once controlling for diachronic patterns of language usage.

Finally, a limitation of the current study that needs to be mentioned involves its scope, both in terms of time span and language. Monaghan and Roberts (2019) provided evidence of cross-linguistic differences in how language change relates to acquisition and word-level properties, even when considering rather similar languages as English and Dutch. Since our study only focused on English, further studies are necessary to probe whether the patterns we uncovered hold cross-linguistically. Moreover, we used the CoHA (Davies, 2010) to derive temporal word embeddings and measures of language change. Whether 200 years of history of a language, characterized using newspaper and magazine articles as well as fiction and nonfiction books, are representative enough to capture the relevant patterns involved in acquisition and evolution is an open question. Our choice was primarily constrained by data availability, since a large-scale corpus is required to reliably train the model we used. Alternative corpora that covered a longer time span, such as Google Books, had other substantial shortcomings related to the availability of rich linguistic context. Future studies are necessary to qualify how the history of a language should be best characterized.

To conclude, we provide evidence of a robust relation between diachronic patterns of language use and *AoA* in English. We introduced and validated measures of language change for a large set of words, derived in an unsupervised way and with no need for annotated data or carefully constructed resources that need expert knowledge. Finally, we showed that language change is best characterized as involving word-level and neighborhood-level patterns, with both relating to *AoA* albeit in different ways.

Acknowledgments

We thank Padraic Monaghan, Thomas Hills, and Andreas Baumann for their insightful reviews of earlier versions of this manuscript. Moreover, GC thanks Max Louwerse for interesting insights during the conceptualization of the study.

Notes

1. We thank Thomas Hills for this remark.
2. As Andreas Baumann pointed out in his review of a first draft of this work, however, it is important to note that Monaghan and Roberts (2019) found a nonlinear effect of frequency on the probability of borrowing, with very low-frequency words being less likely to be borrowed. This suggests that the picture is likely more complex than a simple pressure to conform to the language more popular in the community. We return on this aspect later in the paper, highlighting how frequency could relate to domain specificity, such that low-frequency words that are less likely to be borrowed

may reflect very specific usages, with a low overall frequency but a possibly high in-domain frequency.

3. Even though word embeddings are typically used to encode lexical semantics, we do not refer to our model as tracking semantic change alone, but rather language change at a more general level, as word embeddings reflect usage patterns and also encode structural relations across words (Hewitt & Manning, 2019; Westbury & Hollis, 2019). However, it is worth pointing out that our measures of change do not capture language evolution at all levels: Change in word form and in sublexical structural properties at the level of phonology and morphology are not tracked by our current model.
4. The corpus can be obtained at https://www.corpusdata.org/.
5. There are several ways to compute contextual diversity scores (see Hollis, 2020, and references therein), which differ in what context they consider (from whole documents, Adelman, Brown, & Quesada, 2006, to unique word types occurring in a narrow window around the target word, Hills, 2013; Hills et al., 2010) and in how they encode it, whether counting unique contexts or using distributed representations (see Cevoli, Watkins, & Rastle, 2020, and references therein). Since determining which operationalization of contextual diversity best accounts for acquisition patterns is outside the scope of the current work, we decided to use the standard measure in the literature, that is, the one introduced by Adelman et al. (2006).
6. See https://github.com/stephantul/old20.
7. The source code for the TWEC model can be found at https://github.com/valedica/twec.
8. Superscript symbols describe the temporal slice, while subscript symbols are used to represent an indexed word of the vocabulary.
9. The code implemented to compute measures of semantic change and perform the statistical analyses detailed in Subsection 2.5 is available at the following GitHub repository: https://github.com/GiovanniCassani/semanticShift_AoA.
10. During an exploratory stage, we considered other measures of central tendency and spread, including the median, median absolute deviation, range, and standard deviation, finding that the sum provides the most information.
11. The tool is available at https://spacy.io/.
12. We included all covariates in spite of collinearity because removing them resulted in a worse fit as measured using *AIC*. We do, however, observe adverse effects of collinearity, which manifest in two regression coefficients changing direction. In detail, while length has a positive effect on *AoA* when considered alone, with longer words being learned later, we observe a weak opposite effect once controlling for other predictors. Similarly, frequency has a negative effect when considered alone (more frequent words are learned earlier), but it changes sign once contextual diversity is included, such that frequent words appear to be learned later once controlling for their contextual diversity. The observation that once both contextual diversity and frequency are included in the same model, the former trumps the latter, with a larger effect size and an effect that keeps its original sign, aligns with previous results on the relation between the two variables in several contexts (Adelman et al., 2006).

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*, 814–823.

Baayen, H. R., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, *30*, 1174–1220.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*, 673–721.

Baumann, A. (2018). Linguistic stability increases with population size, but only in stable learning environments. In C. Cuskley, M. Flaherty, H. Little, L. McCrohon, A. Ravignani, & T. Verhoef (Eds.), *The evolution of language: Proceedings of the 12th international conference (EVOLANGXII)*. https://doi.org/10.12775/3991-1.004

Bianchi, F., Di Carlo, V., Nicoli, P., & Palmonari, M. (2020). Compass-aligned distributional embeddings for studying semantic differences across corpora. Available at: https://arxiv.org/abs/2004.06519.

Biemiller, A., Rosenstein, M., Sparks, R., Landauer, T. K., & Foltz, P. W. (2014). Models of vocabulary acquisition: Direct tests and text-derived simulations of vocabulary growth. *Scientific Studies of Reading*, *18*, 130–154.

Bowern, C. (2019). Semantic change and semantic stability: Variation is key. In Proceedings of the 1st international workshop on computational approaches to historical language change (pp. 48–55). Florence, Italy: Association for Computational Linguistics.

Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process*. Chicago, IL: University of Chicago Press.

Bradley, V., Davies, R., Parris, B., Su, I. F., & Weekes, B. S. (2006). Age of acquisition effects on action naming in progressive fluent aphasia. *Brain and Language*, *99*, 128–129.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, *3*, 52–67.

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1–47.

Brysbaert, M. (2017). Age of acquisition ratings score better on criterion validity than frequency trajectory or ratings "corrected" for frequency. *Quarterly Journal of Experimental Psychology*, *70*, 1129–1139.

Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition*, *13*, 992–1011.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*, 991–997.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*, 183–186.

Cevoli, B., Watkins, C., & Rastle, K. (2020). What is semantic diversity and why does it facilitate visual word recognition? *Behavior Research Methods*, 1–17. https://doi.org/10.3758/s13428-020-01440-1.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489–509.

Croft, W. (2000). *Explaining language change: An evolutionary approach*. London: Pearson Education.

Davies, M. (2010). *The corpus of historical American English (CoHA): 400 million words, 1810–2009*. Provo, UT: Brigham Young University.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies (NAACL-HLT) (Vol. *1*, pp. 4171–4186). Stroudsburg, PA: Association for Computational Linguistics.

Di Carlo, V., Bianchi, F., & Palmonari, M. (2019). Training temporal word embeddings with a compass. In Proceedings of the AAAI conference on artificial intelligence (Vol. *33*, pp. 6326–6334). https://doi.org/10.1609/aaai.v33i01.33016326

Dubossarsky, H., Hengchen, S., Tahmasebi, N., & Schlechtweg, D. (2019). Time-out: Temporal referencing for robust modeling of lexical semantic change. In The 57th annual meeting of the Association for Computational Linguistics (ACL2019) proceedings of the conference (pp. 457–470). Florence, Italy: Association for Computational Linguistics.

Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1103.

Firth, J. R. (1957). *Papers in linguistics, 1934–1951*. Oxford, England: Oxford University Press.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*, E3635–E3644.

Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica*, *115*, 43–67.

Gilhooly, K. J., & Gilhooly, M. L. (1980). The validity of age-of-acquisition ratings. *British Journal of Psychology*, *71*, 105–110.

Gonen, H., Jawahar, G., Seddah, D., & Goldberg, Y. (2020). Simple, interpretable and stable method for detecting words with usage change across corpora. In Proceedings of the 58th annual meeting of the Association for Computational Linguistics (pp. 538–555). Stroudsburg, PA: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.acl-main.51.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*, 1006–1033.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016a). Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing (Vol. *2016*, pp. 2116–2121). Stroudsburg, PA: Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016b). Diachronic word embeddings reveal statistical laws of semantic change. In Proceedings of the 54th annual meeting of the Association for Computational Linguistics (pp. 1489–1501). Berlin: Association for Computational Linguistics. Available at: https://www.aclweb.org/anthology/P16-1141

Harris, Z. (1954). Distributional structure. *Word*, *10*, 146–152.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019) (pp. 4129–4138). Stroudsburg, PA: Association for Computational Linguistics.

Hills, T. (2013). The company that words keep: Comparing the statistical structure of child-versus adult-directed language. *Journal of Child Language*, *40*, 586–604.

Hills, T. T., & Adelman, J. S. (2015). Recent evolution of learnability in American English from 1800 to 2000. *Cognition*, *143*, 87–92.

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language*, *63*, 259–273.Available at: https://www.ncbi.nlm.nih.gov/pubmed/20835374

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. B. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science*, *20*, 729–39.Available at: https://www.ncbi.nlm.nih.gov/pubmed/19470123.

Hills, T. T., Proto, E., Sgroi, D., & Seresinhe, C. I. (2019). Historical analysis of national subjective wellbeing using millions of digitized books. *Nature Human Behaviour*, *3*, 1271–1275.

Hilpert, M., & Perek, F. (2015). Meaning change in a petri dish: Constructions, semantic vector spaces, and motion charts. *Linguistics Vanguard*, *1*, 339–350.

Hodgson, C., & Ellis, A. W. (1998). Last in, first to go: Age of acquisition and naming in the elderly. *Brain and Language*, *64*, 146–163.

Hollis, G. (2020). Delineating linguistic contexts, and the validity of context diversity as a measure of a word's contextual variability. *Journal of Memory and Language*, *114*, 104146.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, *23*, 1744–1756.

Holmes, S. J., Jane Fitch, F., & Ellis, A. W. (2006). Age of acquisition affects object recognition and naming in patients with Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology*, *28*, 1010–1022.

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, *131*, 684.

Kanjirangat, V., Mellace, S., & Alessandro, A. (2020). Temporal embeddings and transformer models for narrative text understanding. In Proceedings of Text2Story —Third workshop on narrative extraction from texts co-located with 42nd European conference on information retrieval (ECIR 2020) (pp. 71–77). Available at: arXiv:2003.08811

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 36–45). Stroudsburg, PA: Association for Computational Linguistics.

Kiela, D., & Clark, S. (2015). Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 2461–2470). Stroudsburg, PA: Association for Computational Linguistics.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., & Petrov, S. (2014). Temporal analysis of language through neural language models. In Proceedings of the ACL 2014 workshop on language technologies and computational social science (pp. 61–65). Stroudsburg, PA: Association for Computational Linguistics.

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*, 102–110.

Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically significant detection of linguistic change. In Proceedings of the 24th international conference on World Wide Web (pp. 625–635). New York: ACM.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*, 978–990.

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. In Proceedings of the 27th international conference on computational linguistics (pp. 1384–1397). Stroudsburg, PA: Association for Computational Linguistics.

Labov, W. (2001). *Principles of linguistic change Volume 2: Social factors, Language in society (Vol. 29)*. Oxford, England: Blackwell.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151–171.

Li, Y., Engelthaler, T., Siew, C. S., & Hills, T. T. (2019). The macroscope: A tool for examining the historical structure of language. *Behavior Research Methods*, *51*, 1864–1877.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*, 713–716.

Lopopolo, A., & van Miltenburg, E. (2015). Sound-based distributional models. In Proceedings of the 11th International Conference on Computational Semantics (pp. 70–75). Stroudsburg, PA: Association for Computational Linguistics.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*, 203–208.

Łuniewska, M., Haman, E., Armon-Lotem, S., Etenkowski, B., Southwood, F., Anđelković, D., Blom, E., Boerma, T., Chiat, S., de Abreu, P. E., Gagarina, N., Gavarró, A., Håkansson, G., Hickey, T., de López, K. J., Marinis, T., Popovićc, M., Thordardottir, E., Blažienė, A., Cantú Sánchez, M., Dabašinskienė, I., Ege, P., Ehret, I.-A., Fritsche, N.-A., Gatt, D., Janssen, B., Kambanaros, M., Kapalková, S., Kronqvist, B., Kunnari, S., Levorato,

C., Nenonen, O., Fhlannchadha, S. N., O'Toole, C., Polišenská, K., Pomiechowska, B., Ringblom, N., Rinker, T., Roch, M., Savić, M., Slančová, D., Tsimpli, I. M., & Ünal-Logacev, Ö. (2016). Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior Research Methods*, *48*, 1154–1177.

MacNeilage, P. F., & Davis, B. L. (2000). On the origin of internal structure of word forms. *Science*, *288*, 527–531.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111–3119). Available at: arXiv:1310.4546

Monaghan, P. (2014). Age of acquisition predicts rate of lexical evolution. *Cognition*, *133*, 530–534.

Monaghan, P., & Roberts, S. G. (2019). Cognitive influences in language evolution: Psycholinguistic predictors of loan word borrowing. *Cognition*, *186*, 147–158.

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *Quarterly Journal of Experimental Psychology Section A*, *50*, 528–559.

Newberry, M. G., Ahern, C. A., Clark, R., & Plotkin, J. B. (2017). Detecting evolutionary forces in language change. *Nature*, *551*, 223–226.

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, *449*, 717–720.

Perek, F. (2014). Vector spaces for historical linguistics: Using distributional semantics to study syntactic productivity in diachrony. In Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA: Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In Proceedings of NAACL-HLT (pp. 2227–2237). Stroudsburg, PA: Association for Computational Linguistics.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Rinaldi, L., & Marelli, M. (2019). The use of number words in natural language obeys Weber's law. *Journal of Experimental Psychology: General*, *149*, 1215–1230.

Rudolph, M., & Blei, D. (2018). Dynamic embeddings for language evolution. In Proceedings of the 2018 World Wide Web conference (pp. 1003–1011). New York: ACM.

Sagi, E., Kaufmann, S., & Clark, B. (2011). Tracing semantic change with latent semantic analysis. *Current Methods in Historical Semantics*, *73*, 161–183.

Schakel, A. M., & Wilson, B. J. (2015). Measuring word significance using distributed representations of words. Available at: arXiv:1508.02297

Smith, K. (2004). The evolution of vocabulary. *Journal of Theoretical Biology*, *228*, 127–142.

Soni, S., Klein, L. F., & Eisenstein, J. (2021). Abolitionist networks: Modeling language change in nineteenth-century activist newspapers. *Journal of Cultural Analytics*, *1*, 1–43.

Steels, L. (2017). Human language is a culturally evolving system. *Psychonomic Bulletin & Review*, *24*, 190–193.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*, 41–78.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: With special reference to north American Indians and Eskimos. *Proceedings of the American Philosophical Society*, *96*, 452–463.

Tahmasebi, N., Borin, L., & Jatowt, A. (2018). Survey of computational approaches to lexical semantic change. Available at: arXiv:1811.06278

Tahmasebi, N., Borin, L., Jatowt, A., & Xu, Y. (Eds.) (2019). *Proceedings of the 1st international workshop on computational approaches to historical language change*. Florence, Italy: Association for Computational Linguistics. Available at: https://www.aclweb.org/anthology/W19-4700

Thomason, S. G., & Kaufman, T. (1992). *Language contact, creolization, and genetic linguistics*. Berkeley, CA: University of California Press.

Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences*, *113*, 4530–4535.Available at: https://www.ncbi.nlm.nih.gov/pubmed/27044094.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Vejdemo, S., & Hörberg, T. (2016). Semantic factors predict the rate of lexical replacement of content words. *PLoS One*, *11*, e0147924.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th edn.). New York: Springer. Available at: http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0

Westbury, C., & Hollis, G. (2019). Conceptualizing syntactic categories as semantic categories: Unifying part-of-speech identification and semantics using co-occurrence vector averaging. *Behavior Research Methods*, *51*, 1371–1398.

Winter, B., Thompson, G., & Urban, M. (2014). Cognitive factors motivating the evolution of word meanings: Evidence from corpora, behavioral data and encyclopedic network structure. In Evolution of Language: Proceedings of the 10th International Conference (EVOLANG10) (pp. 353–360). Singapore: World Scientific.

Wittgenstein, L. (1953). *Philosophical investigations (Translated by Anscombe, G. E. M.)*. Oxford, England: Basil Blackwell.

Yao, Z., Sun, Y., Ding, W., Rao, N., & Xiong, H. (2018). Dynamic word embeddings for evolving semantic discovery. In Proceedings of the eleventh ACM international conference on web search and data mining (pp. 673–681). New York: ACM.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*, 971–979.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Boston, MA: Addison-Wesley Press.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting information