Coherent oscillations in word-use data from 1700 to 2008

palgrave communications
HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

Received 10 May 2016 | Accepted 1 Nov 2016 | Published 22 Nov 2016

DOI: 10.1057/palcomms.2016.84 OPEN

Coherent oscillations in word-use data from 1700 to 2008

Marcelo A Montemurro[1] and Damián H Zanette[2]

ABSTRACT In written language, the choice of specific words is constrained by both grammatical requirements and the specific semantic context of the message to be transmitted. To a significant degree, the semantic context is in turn affected by a broad cultural and historical environment, which also influences matters of style and manners. Over time, those environmental factors leave an imprint in the statistics of language use, with some words becoming more common and other words being preferred less. Here we characterize the patterns of language use over time based on word statistics extracted from more than 4.5 million books written over a period of 308 years. We find evidence of novel systematic oscillatory patterns in word use with a consistent period narrowly distributed around 14 years. The specific phase relationships between different words show structure at two independent levels: first, there is a weak global phase modulation that is primarily linked to overall shifts in the vocabulary across time; and second, a stronger component dependent on well defined semantic relationships between words. In particular, complex network analysis reveals that semantically related words show strong phase coherence. Ultimately, these previously unknown patterns in the statistics of language may be a consequence of changes in the cultural framework that influences the thematic focus of writers.

[1] The University of Manchester, Faculty Biology, Medicine and Health, Manchester, UK [2] Centro Atómico Bariloche and Instituto Balseiro, Comisión Nacional de Energía Atómica. Consejo Nacional de Investigaciones Cientficas y Técnicas, Río Negro, Argentina Correspondence: (e-mail: m.montemurro@manchester.ac.uk)

Introduction

The application of quantitative methods to the analysis of language has disclosed layers of non-trivial statistical structure ranging from word frequencies (Ferrer-i-Cancho and Solé, 2001a; Montemurro, 2001), to correlations between hundreds or thousands of words (Montemurro and Pury, 2002; Alvarez-Lacalle *et al.*, 2006), and to the complex network organization of the whole lexicon (Ferrer-i-Cancho and Solé, 2001b; Sigman and Cecchi, 2002). Likewise, the statistical structure of large collections of texts revealed novel quantitative linguistic universals that depend on the ordering of words (Montemurro and Zanette, 2011). Far from being static, language is a dynamic entity showing patterns of change over time that range from the scale of a few years, as in a wave of fashion, to episodes of birth and death of language families that may take place over many thousand years. The study of these processes suggested that language can be understood as an evolutionary system (Nowak *et al.*, 2002) bearing strong similarities to mechanisms underlying the evolution of biological species (Pagel, 2009), which had been originally recognized by Charles Darwin (Darwin, 1871). The application of phylogenetic methods originally devised to characterize the evolution of biological organisms has yielded insight into the transformation of languages at the macro-scale of thousands of years (Gray and Atkinson, 2003), allowing to make inferences about language evolution reaching back to the the Upper-Palaeolithic period (Pagel *et al.*, 2013). On shorter time scales, the analysis of grammatical and morphological changes has shed light on the dynamics of language over the course of half a millennium (Lieberman *et al.*, 2007).

The availability of large amounts of linguistic data after the expansion of the Internet opened a new era for quantitative studies in language dynamics. The analysis of large volumes of digitized text provided insights into the statistics of word use in social communication groups (Altmann *et al.*, 2011), changes in literary style over a historical corpus of literary works (Hughes *et al.*, 2012), and sentiment analysis in Twitter (Twenge *et al.*, 2012), among others.

In a broader context, language usage can be regarded as an important component among the factors that contribute to complex interactions within social systems (Castellano *et al.*, 2009). As such, language is both an input and an output within and across groups in human societies. Whereas notable contingencies like important historical events can exert a strong influence on subsets of the vocabulary used at a particular time, slower processes of purely linguistic nature or cultural change can manifest themselves as overall trends in the use of certain words. Until recently, the study of this type of patterns in language evolution had to rely on scattered and sparse evidence from a reduced number of written sources. This situation suddenly changed in 2010 when Google Inc. made available the Google Ngram database, consisting of word frequency counts from nearly 5 million digitized books covering a range of more than 500 years. The initial studies carried out on this large database have allowed scholars to address unprecedented questions about language usage. In particular, for the first time it was possible to study quantitatively aspects of cultural change as reflected in language (Michel *et al.*, 2011; Greenfield, 2013), and rigorously assess overall vocabulary drift over the time span of two centuries (Bochkarev *et al.*, 2014). Moreover, methods inspired in statistical mechanics of complex systems were used to study the dynamics of word birth and death (Petersen *et al.*, 2012a), long-range fractal correlations in word frequencies over centuries (Gao *et al.*, 2012), and the scaling behaviour of word frequencies over time as represented by Zipf's (1949) and Heaps' (1978) laws (Petersen *et al.*, 2012b; Gerlach and Altmann, 2013).

In the present study we employed data from the Google Ngram database to analyse the temporal evolution in the use of words over the span of more than three centuries, from 1700 to 2008. In particular, we focused on systematic patterns of change in the relative prevalence of common nouns over time. The noun class was chosen because we expected it to be the most semantically relevant group of words and, as such, to bear more information on patterns of cultural change over the time scales considered. However, we also obtained results for verbs which, although affected to a lesser degree, also showed systematic variation over time.

As a natural hierarchical measure of the prevalence of each individual word i within a given set of words at time t, we propose to consider the word's rank $r_i(t)$ in a list where the set is ordered by decreasing frequencies. This measure allows to quantify the relative changes in importance—or popularity—among words along time (Cocho *et al.*, 2015). An additional reason for focusing on the rank is that the instantaneous frequency of words are, in fact, strongly dependent on corpus size—which, in our case, grows by orders of magnitude from 1700 to 2008. For instance, the English corpus of 1-grams—that is single words—changes from a total of 3.7 million words in 1700 to more than 19 billions in 2008. The rank, on the other hand, is limited to positive integers up to the lexicon size, and is therefore expected to grant a more robust measure of word relevance (see Supplementary Information for a detailed analysis of this point).

Our analysis reveals the presence of persistent oscillations in word use, shared by all the words studied and with a well defined period of around 14 years. Moreover, these oscillations present a rich relationship in their phases, in the form of two largely independent features: one consists of a global word-independent phase modulation related to overall shifts in the vocabulary at specific times, while the other is given by a word-dependent modulation that induces coherent phase behaviour for semantically related groups of words.

To quantify the relative change in the hierarchical position of a word, we introduce the following quantity:

$$\rho_i(t) = \ln r_i(t) - \ln r_i(t+1), \tag{1}$$

where the time t is measured in years. This quantity gives the logarithmic rank variation per year, and is closely related to the log frequency return used by Petersen *et al.* (2012a). For any given word i, $\rho_i(t)$ represents a time series quantifying changes in relative word prevalence. Our results concern the ensemble information present in the temporal behaviour given by Equation (1) for all i's. The details of our analysis are presented for the English language, but evidence of similar behaviour is provided for French, German, Italian, Russian and Spanish.

The specific data for our study come from the 2012 version of the Google Ngram database. We focus on 1-grams data, which consist of single word frequency counts for every year independently. The 2012 version of the database is annotated with parts-of-speech tags (Lin *et al.*, 2012), which allow the extraction of particular word classes. In all cases, we considered common nouns, defined as those that appeared at least 50,000 times over the whole period considered. The rank for each noun in each year was defined from the subset of extracted common nouns. Capitalization was ignored and, in the case of English, all nouns were converted to singular form in order to increase the statistics. Finally, Equation (1) requires that the rank is defined for all years considered. Thus, we only kept a core vocabulary of 5,630 common nouns that were used in every year from 1700 to 2008.

Methods

Data preparation. We used the 2012 version of the Google Ngram database for 1-grams, consisting of word frequency counts per year in the interval 1520–2008 (Google Inc., 2013). For English, our starting database was the 1-gram counts from 1700 because, although the available data start at 1520, the data for the first 200 years is rather sparse. Using the parts-of-speech tag included in the database (Lin *et al.*, 2012), we extracted the 1-gram information for the noun and verb classes. In order to increase statistics, all nouns were converted to singular form and verbs to their infinitive forms. From these words we only kept those that had an accumulated occurrence of at least 50,000 times in the interval 1700–2008 and that had been used in every single year in that interval. A further restriction was to use words that were written only using letter characters—thus avoiding numbers and other special characters. This procedure finally left a core vocabulary of 5,360 nouns and 2,342 verbs. For other languages, because of higher data sparseness in the 18th century, we carried out the same procedure starting from 1800 (see Supplementary Table S1).

Wavelet analysis and pseudo-period. The wavelet transform of a real function $g(t)$ is defined as

$$w(u,\ s) = \frac{1}{\sqrt{s}}\int_{-\infty}^{\infty} g(t)\psi\left(\frac{t-u}{s}\right)dt, \tag{2}$$

where u is the time shift and s is the wavelet scale. In our analysis we use the Mexican Hat Wavelet, given by

$$\psi(x) = \frac{2}{(9\sigma^2\pi)^{1/4}}\left(1-\frac{x^2}{\sigma^2}\right)\exp\left(-\frac{x^2}{2\sigma^2}\right), \tag{3}$$

with $\sigma = 1$. A natural way to determine a period T corresponding to the scale s is to find the extrema over s of the wavelet transform $w(u, s)$ of a periodic function $g(t) = \cos(2\pi t/T)$. By computing the transform using the kernel defined in Equation (3), the period that maximizes the absolute value of the wavelet coefficients at scale s is given by $T = (8\pi^2/5)^{1/2}s$.

Clustering and network structure. In order to apply the clustering algorithm, every word i was represented by the time series $\rho_i(t)$, and the correlation between any two time series was used to quantify similarity between the corresponding words. The correlation between $\rho_i(t)$ and $\rho_i(t)$ is defined as

$$C(i,\ j) = \frac{\left\langle \rho_i(t)\rho_j(t)\right\rangle - \left\langle \rho_i\,(t)\right\rangle\left\langle \rho_j(t)\right\rangle}{\sigma_i\sigma_j} \tag{4}$$

where the averages are taken over time, and $\sigma_{i,j}$ represent the respective standard deviations. Then, a distance can be defined between words i and j as $D(i, j) = 1 - C(i, j)$. The clustering algorithm proceeds by progressive agglomeration. It starts assuming as many initial clusters as words, and then groups the pair of words having the closest distance. It then proceeds iteratively, merging the closest clusters into larger ones. In our particular implementation, the distance between two clusters was taken as the average of all the distances between the elements belonging to the two clusters.

The word network is built by establishing links between words whose correlation equals or exceeds a given threshold θ. Thus, the corresponding adjacency matrix is defined as

$$A_{ij} = \begin{cases} 1 & \text{if} \quad C(i,\ j) \geqslant \theta\ , \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The division of the networks into communities is based on the maximization of network modularity (Clauset *et al.*, 2004; Newman, 2006), which is defined as

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i,\ c_j), \tag{6}$$

with m the total number of links in the network. In this expression, the sum runs over all the pairs of nodes $\{i, j\}$, and $k_{i,j}$ are the respective degrees. The Kronecker symbol $\delta(c_i, c_j)$ equals 1 if i and j belong to the same community, and 0 otherwise.

Results

As is the case of frequency, the rank of words undergo changes year on year, reflecting their relative prevalence in usage (Cocho *et al.*, 2015). Fig. 1a and c show the evolution of the rank as a function of time for two groups of semantically related words. The selected words occupy a wide range of rank positions. While, for instance, the word *king* fluctuates over relatively low ranks (corresponding to high frequencies), *duchess* occupies always a rank higher than 1,000. A similar situation is found for the food related terms, where the ranks of *food* and *chicken* differ from each other by some thousands. As for their time variations, although there is some correspondence in the positions of local maxima and minima within each group, correlations over longer time spans are generally weak.

However, as shown in Fig. 1b and d, a coherent oscillatory pattern is revealed when we look at the logarithmic rank variation $\rho_i(t)$, with different curves in each group following closely similar behaviour. All the words in each group show a remarkably consistent common pattern, in which they systematically increase their popularity over certain intervals and decline in the intervening years. In the following, we quantify this observation at levels ranging from individual words to semantically related groups.

Periods and phases of oscillations. To characterize the periodicity of the oscillations we first estimated the periods of the individual words by means of a wavelet analysis of their respective ρ_i time series. Figure 2 shows two examples of the procedure

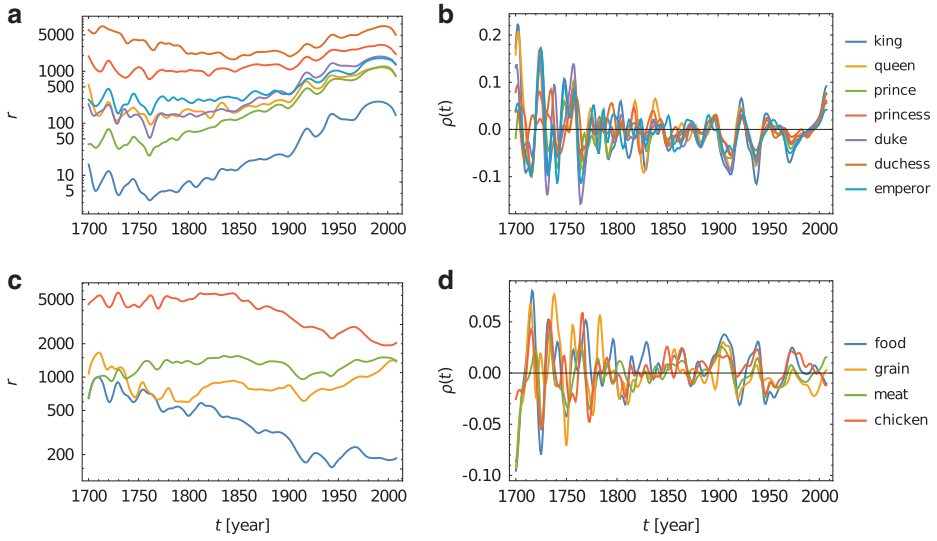

Figure 1 | Time evolution of ranks. (a, c) Variations of the logarithm of the rank over three centuries for two groups of related words. (b, d) Yearly differences in the logarithm of the rank of the same words, reflecting oscillatory variations in word prevalence.

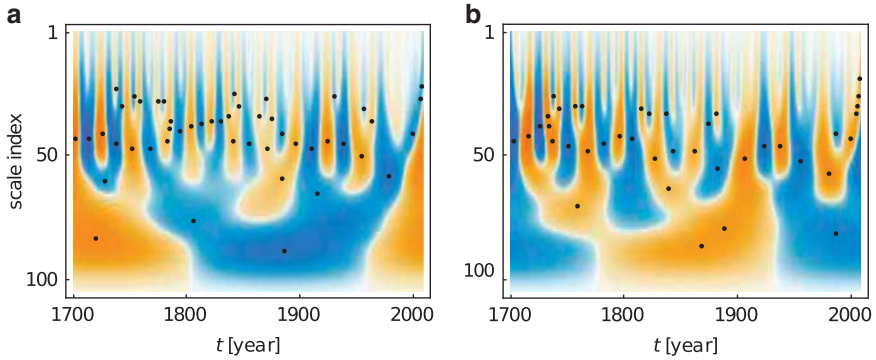

Figure 2 | Example scalograms showing local extrema of wavelet coefficients. Colours towards yellows and blues indicate positive and negative values, respectively. Dots indicate the position of the local maxima and minima from which the period is estimated as explained in Methods. (a) Scalogram for *king*; (b) scalogram for *food*.

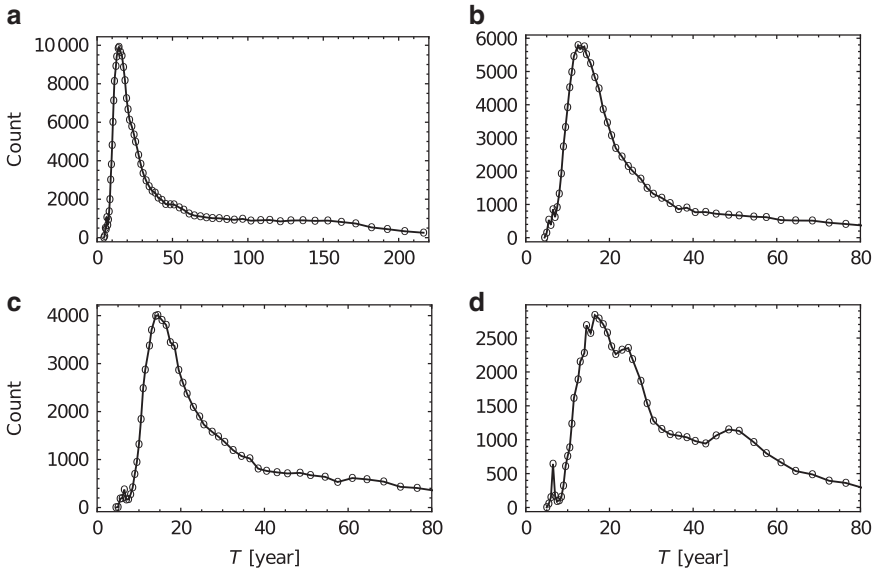

Figure 3 | Period of oscillations. Histograms showing the distribution of oscillation periods obtained from the wavelet analysis explained in the text, (a) for all years since 1700; (b) for the 18th century; (c) for the 19th century; (d) for the 20th century.

we used to obtain the periods for every word in the core vocabulary. Briefly, we computed a scalogram and then obtained the set of local extrema. Each of these extrema can be associated to a pseudo-period, from which the histogram of periods is then computed (see Methods for details).

Figure 3 shows the resulting distribution of periods, for the whole time range (Fig. 3a), and discriminated per century (Fig. 3b to d). Fig. 3a shows a narrow peak at around 14 years with a small kink close to 50 years. In the figures corresponding to the individual centuries, it is apparent that oscillatory modes with longer periods increase in importance from century to century. In particular, the kink around 50 years is clearly a contribution of the 20th century. The effect can also be noticed in the individual time series of Fig. 1b and d as a tendency of the oscillations to slow down towards the present.

In addition to the period of signals, another aspect related to the specific timing structure of the oscillations is given by their phase. While the data in Fig. 1b and d show that the two groups of words exhibit similar oscillation periods, the phases between them are different. For instance, towards the year 1900 the first group is changing downwards while the second group follows an opposite trend.

The study of phase relationships across the whole core vocabulary reveals two independent modulations affecting the phase of oscillations. Figure 4a shows the time evolution of all 5,360 nouns arranged in a matrix-like structure following a random order. Yellow and blue shades respectively indicate high (positive) and low (negative) values of $\rho_i(t)$. As put in evidence by the vertical strips of either yellowish or bluish tonalities, the presence of a global modulation in the phase, affecting all words more or less uniformly, is apparent. There are specific time ranges in which the words in the core vocabulary move preferably towards higher ranks, while in other intervals they tend to move down. These events signal major shifts in the overall lexicon: the fact that all nouns in the core vocabulary move towards higher ranks means that other nouns, which are not part of the core, become temporarily more important. It is striking that these events occur repeatedly with effects all across the core nouns. The curve on top of the matrix is the average of $\rho_i(t)$ over the core vocabulary, representing the mean modulation of variations in its usage.

Clusters and networks of nouns. To analyse relationships in the time evolution that are more dependent on specific words, the

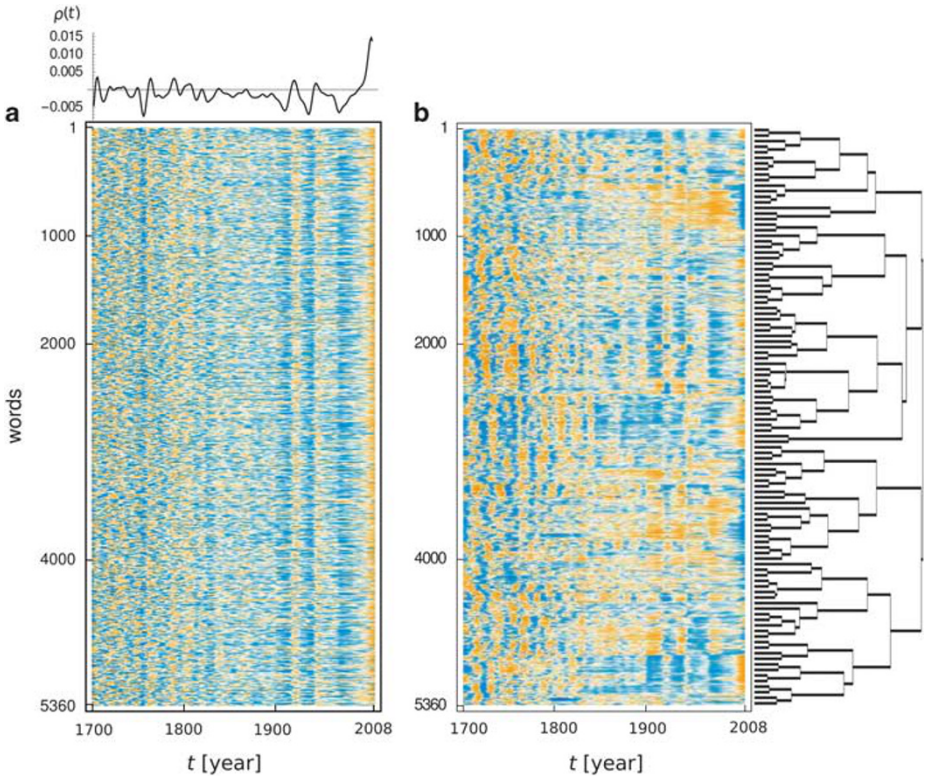

Figure 4 | Phase relationships. (a) Top: Average time evolution of ρ_i for the 5,360 nouns in the core vocabulary. Bottom: Individual evolution of ρ_i for the same words, arranged in random order. Yellow and blue shades indicate high (positive) and low (negative) values. (b) Left: Individual evolution of ρ_i after word reordering according to the results of a hierarchical clustering algorithm. Right: Tree structure corresponding to the first few levels of clustering.

mean modulation over the core vocabulary was subtracted from the time evolution of all words. The resulting time series were then grouped by means of a hierarchical clustering algorithm (see Methods for details) leading to a hierarchical tree structure, where closer topological distance across the dendrogram means greater similarity in the time series of the words. The first levels of the resulting tree are depicted in Fig. 4b together with the corresponding reordering of the time series. The most remarkable feature in the reordered dataset is the presence of numerous word groups that share specific phase relationship patterns over time. For many of the words, the time evolution is similar to the general trend observed in Fig. 4a, while others exhibit very different behaviours.

By inspecting the sequence of words in the ordering given by the clustering, it is apparent that most of the structure in the dataset is given by groups of semantically related nouns. Table 1 shows a few examples of contiguous groups extracted from the ordering produced by clustering (Fig. 4b). The clear semantic relationship between the words in each group emphasizes the close connection between meaning and changes of relative prevalence in the vocabulary.

While clustering highlights the existence of well defined groups of words with similar time behaviour—and, additionally, with close semantic relationships—the capture of more complex structure within each group requires characterizing detailed relations between word pairs. In Fig. 5a we show the correlation matrix for the time evolution of all the words in the core vocabulary (see Methods). The ordering of indices in the matrix is the same as for the result of clustering. As a consequence, most positive correlations (yellow shades) are distributed along the diagonal. However, the still very significant off-diagonal structure suggests that a description in terms of network topology would be more appropriate to represent the relationships between words. To extract a network from the correlation matrix we introduce a threshold θ. Two words i and j will be connected in the network is their correlation $C(i, j)$ is larger than or equal to the threshold.

For a given value of θ the resulting network is generally not connected, but instead consists of a number of mutually disconnected components. Figure 5b shows the fraction of nodes (that is, words) in the largest component as a function of the threshold. As expected, for sufficiently small values of θ the largest component is comparable in size to the total network, containing a fraction of nodes equal or very close to one. As the threshold grows, however, there is a narrow range around $\theta \approx 0.65$ where the fraction of nodes in the largest component drops rather abruptly, indicating that the network splits into a large number of small components. We verified that at the critical threshold $\theta^* = 0.65$ the degree distribution of the network, shown in Fig. 5c, approximately follows a power law, suggesting scale-free structure (Barabási and Albert, 1999). Figure 5d shows a diagram of the largest connected component at θ^*, comprising 2,670 nodes.

The largest component of the noun network also presents small-world features (Watts and Strogatz, 1998). At the critical threshold, its mean topological distance (diameter) is 6.77, while its mean clustering coefficient is 0.29. These values have to be compared with those obtained for a random graph. We have found that, in the noun network, the largest component has a diameter only 50% larger than the corresponding random graph while, on the other hand, the clustering coefficient is 100 times that of the random counterpart. These values indicate a small-world structure, with both locally strong connectivity and long-range connections joining distant parts of the network.

Table 1 | Semantic affinity of clustered nouns

Index	*Noun*	*Index*	*Noun*	*Index*	*Noun*
349	hannibal	2252	poet	2735	holiness
350	scipio	2253	poem	2736	mercy
351	carthage	2254	poetry	2737	temptation
352	carthaginian	2255	prose	2738	wrath
353	cato	2256	dryden	...	...
354	cicero	2257	muse	4547	crop
355	pompey	2258	ode	4548	wheat
356	marius	2259	verse	4549	husbandry
357	brutus	2260	pastoral	4550	hay
358	legion	2261	virgil	4551	oat
359	consul	2262	ovid	4552	sow
360	tribune	2263	criticism	4553	cow
361	senate	2265	hero	4554	hog
362	senator	2266	imitation	4555	calve
...	...	2267	diction	4556	rot
682	child	...	...	4557	swine
683	parent	2721	god	4558	tree
684	infant	2722	faith	4559	fruit
685	adult	2723	christ	4560	season
686	woman	2724	scripture	4561	winter
687	sex	2725	salvation	4562	summer
688	family	2726	sanctification	4563	cattle
...	...	2727	righteousness	4564	sheep
1241	ship	2728	commandment	4565	pasture
1242	sail	2729	doctrine	4566	grass
1243	anchor	2730	gospel	4567	meadow
1244	voyage	2731	apostle	4568	moss
1245	pirate	2732	worship	4569	gravel
1246	beard	2733	sin	4570	frost
...	...	2734	sinner	4571	soil

The table shows a few examples of consecutive words as reordered by the hierarchical clustering algorithm.

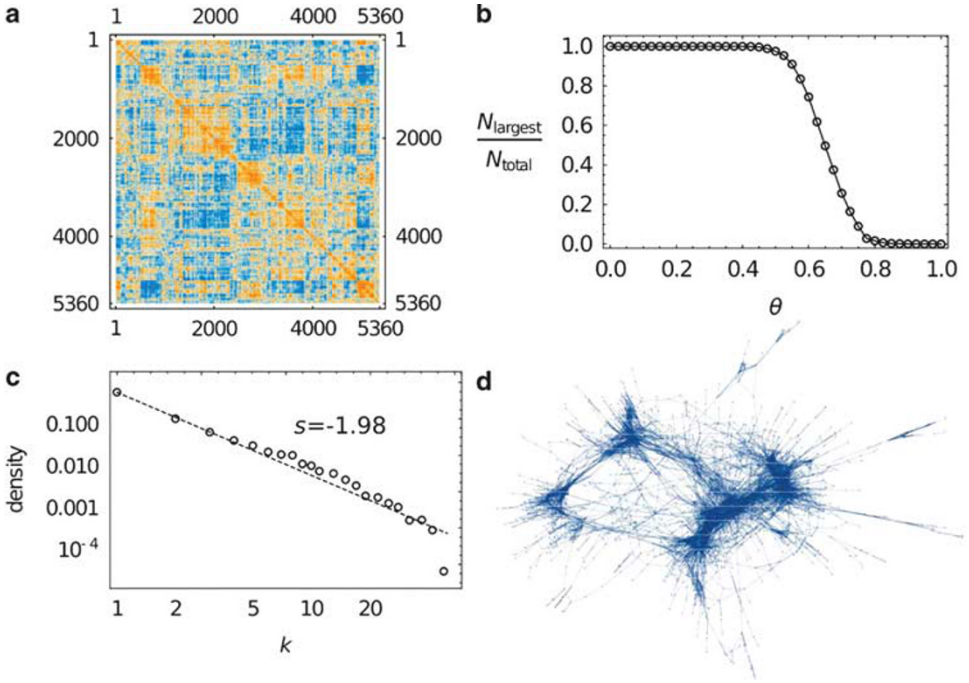

Figure 5 | Noun network structure. (a) Correlation matrix for all words in the core vocabulary. The ordering of indices is the same as results from clustering (see Fig. 4b). Yellow and blue shades indicate high (positive) and low (negative) values. (b) Fraction of nodes in the largest network component as a function of the correlation threshold θ used to build the network. At the transition, around the critical value $\theta^* = 0.65$, the network breaks down into a large number of small components. (c) Network degree distribution at the critical value of the threshold. The approximated power-law profile, with slope $s \approx -2$, is a signature to a scale-free network. (d) Diagram of the largest network component at the critical threshold.

Networks that share these features may have a structure where different parts of the network are naturally segregated into communities. Within each community, nodes are preferentially connected to other members of the same community, with only a few links directed towards other communities. To test this possibility on the noun network, we extracted its communities using a modularity optimization algorithm (Clauset *et al.*, 2004; Newman, 2006) (see Methods) as implemented in Mathematica (Wolfram Research, 2016).

Figures 6 and 7 show two examples of the communities obtained from the noun network. To reveal more structure within each community, we proceeded to further divide them into sub-communities, by simply iterating once more the same algorithm. The community shown in Fig. 6 exhibits a strong

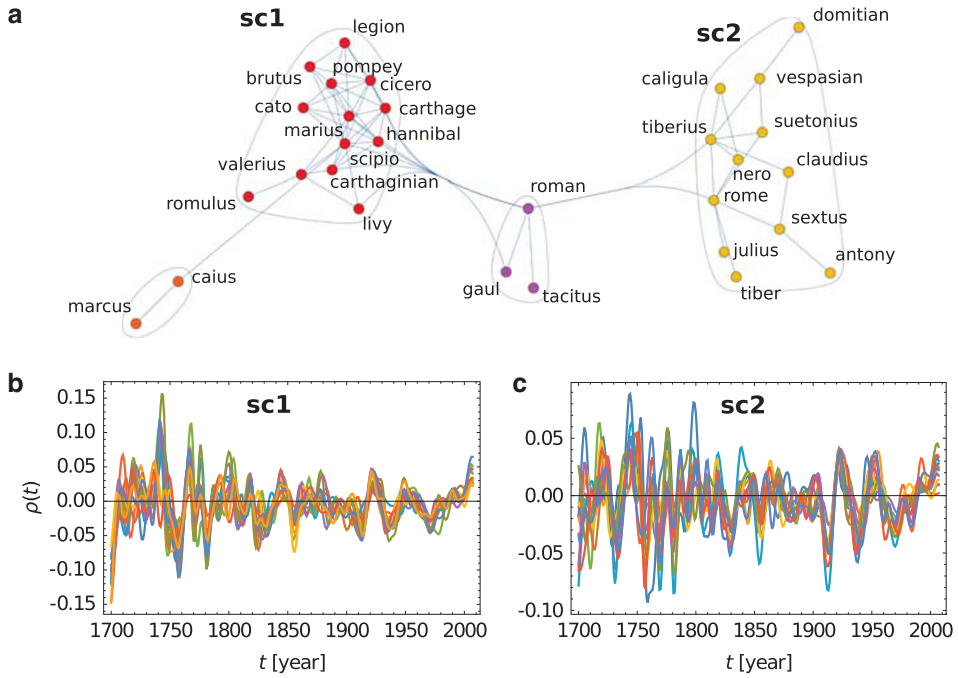

Figure 6 | Example of a community in the noun network. In this case, it contains words related to Ancient Rome. The two largest sub-communities (sc1 and sc2) are readily associated with different Roman historical periods. The lower panels show the evolution of ρ_i for words in different sub-communities.

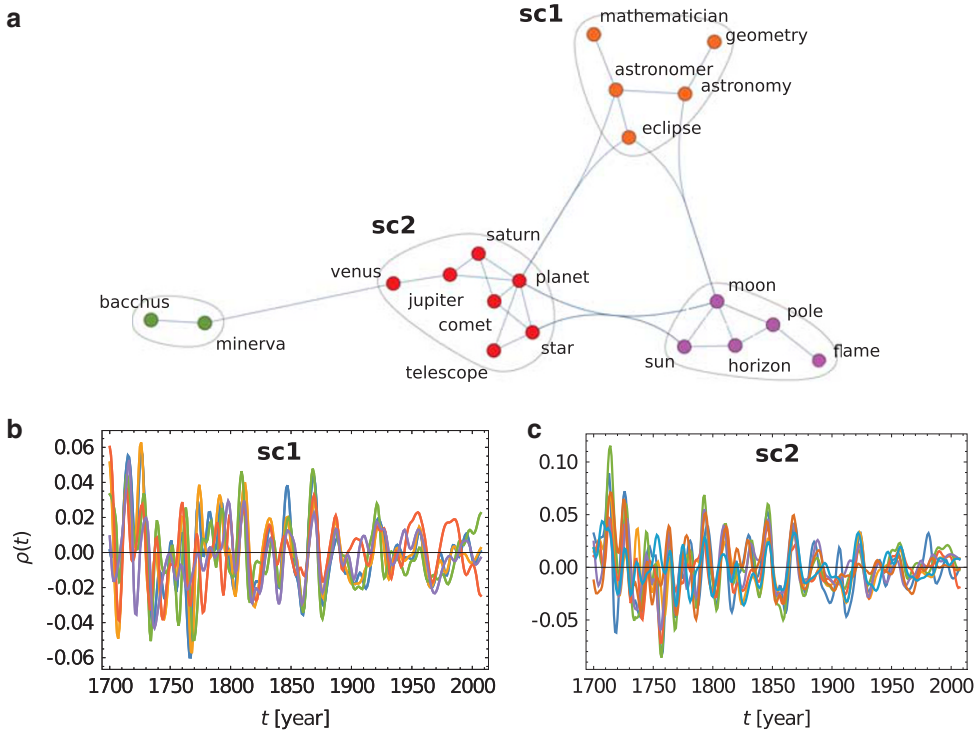

Figure 7 | Example of a community in the noun network. As in Fig. 6, for words related to astronomy.

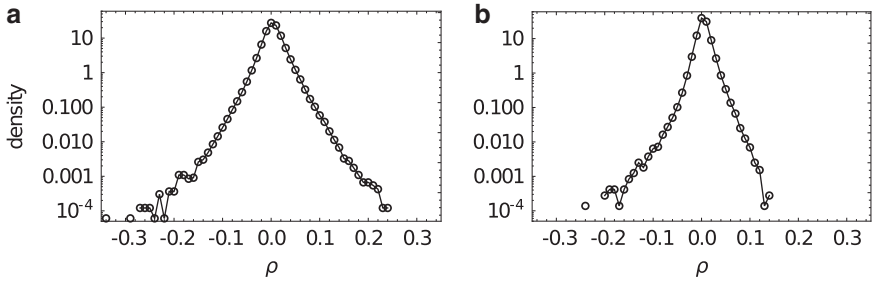

Figure 8 | Distribution of instantaneous amplitudes for the pooled word time-series. (a) Nouns; (b) verbs.

thematic link, consisting almost exclusively of nouns relating to Ancient Rome. Interestingly, the further subdivision into sub-communities shows another layer of structure implicit in the correlations between words. Particularly, the sub-community in which nodes are represented by red disks has a clear pre-Imperial flavour, while the sub-network that consists of yellow disks is strongly Imperial. As the panels including the time series for the members of those two sub-communities show, oscillations have distinct temporal features, similar for words within the same sub-community and different across sub-communities. The community shown in Fig. 7, in turn, is thematically linked with astronomy. As noted above, further subdivision into sub-communities clearly shows that time evolution is strongly correlated for words with strong semantic relationship.

The focus of our analysis has been on the noun class of words, since it represents the most semantically informative group of words. However, we have also checked that similar oscillations are found among verbs, albeit with typically smaller amplitudes compared to those found for nouns. Figure 8 shows histograms obtained from pooling all the values of $\rho_i(t)$ for all nouns and verbs in each dataset.

We verified that oscillatory patterns similar to those found for English are also observed for nouns in French, German, Italian, Russian and Spanish. For these languages, however, the analysis was limited to the last two centuries because of their scarce representation in the Google database during the 1700s. The respective distributions of oscillation periods, in particular, are strikingly coincident with each other (see Supplementary Information).

Discussion

Using data extracted from the Google Ngram database, we have disclosed a systematic oscillatory pattern in the use of common nouns over the last three centuries. This regularity, which has been quantified in terms of the relative prevalence of different words in the vocabulary, was consistently confirmed to occur in a set of several thousand nouns, and was also observed for verbs. Characterization of the oscillations revealed a well-defined period of around 14 years, with a tendency to become longer and more spread towards the 20th century. The phase of oscillations, on the other hand, can broadly vary from word to word, but high correlation between phases was a typical signature of semantic affinity between the respective words. This trend made it possible, on the basis of comparing the oscillations of different words, to build a network whose communities contain nouns related by their meaning.

A preliminary analysis of other languages of the Indo-European family revealed oscillations with similar characteristics. Specifically, the distributions of oscillation periods over the last two centuries were closely similar to that found for English.

At present, we do not have an explanation for the oscillatory behaviour of word prevalence. As advanced in the Introduction, however, we expect that this behaviour is related to changes in the cultural environment that, in turn, stir the thematic focus of the writers represented in the Google database. Oscillatory dynamics, moreover, have been demonstrated in other areas of social sciences, such as in economics (Morgan, 1990), where the quantification of cyclic behaviour is more direct than for cultural changes.

On the other hand, the inference of cultural evolution features from the analysis of Google Ngrams time series has recently been criticised mainly on the basis that a database built from book digitization may be strongly biased towards certain thematic areas, or by a handful of influential writers (Pechenick *et al.*, 2015). However, although this warning is conceptually well grounded, objective evidence that the biases observed in the Google database *do not* respond to genuine trends in cultural focus has not been produced yet. Leaving aside certain variations ascribable to linguistic or stylistic evolution, reported biases correspond either to localized (inter-decade) frequency changes of a few significant words, or to long-range, quasi-monotonic thematic drifts—in particular, towards science and technology. None of these biases, nor their putative causes, point to the possibility of oscillatory behaviour in word usage along the database. In contrast, our observation of sustained oscillations in word prevalence, and the fact that they consistently occur over vocabularies comprising thousands of words in different languages, confer statistical significance to our results, beyond presumptive distortions in the Google three-century-long selection of books.

References

Altmann EG, Pierrehumbert JB and Motter AE (2011) Niche as a determinant of word fate in online groups. *PLoS ONE*; **6** (5): e19009.

Alvarez-Lacalle E, Dorow B, Eckmann JP and Moses E (2006)), Hierarchical structures induce long-range dynamical correlations in written texts. *Proceedings of the National Academy of Sciences*; **103** (21): 7956–7961.

Barabási A-L and Albert R (1999) Emergence of Scaling in Random Networks. *Science*; **286**, 509–512.

Bochkarev V, Solovyev V and Wichmann S (2014) Universals versus historical contingencies in lexical evolution. *Journal of The Royal Society Interface*; **11** (101): 841.

Castellano C, Fortunato S and Loreto V (2009) Statistical physics of social dynamics. *Reviews of Modern Physics*; **81** (2): 591.

Clauset A, Newman ME and Moore C (2004) Finding community structure in very large networks. *Physical review E*; **70** (6): 066111.

Cocho G, Flores J, Gershenson C, Pineda C and Sánchez S (2015) Rank diversity of languages: Generic behavior in computational linguistics. *PLoS ONE*; **10** (4): e0121898.

Darwin C (1871) *The descent of Man, and Selection in Relation to Sex*. J. Murray: London.

Ferrer-i-Cancho R and Solé RV (2001a) Two regimes in the frequency of words and the origins of complex Lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*; **8** (3): 165–173.

Ferrer-i-Cancho R and Solé RV (2001b) The small world of human language. *Proceedings of the Royal Society B: Biological Sciences*; **268** (1482): 2261–2265.

Gao J, Hu J, Mao X and Perc M (2012) Culturomics meets random fractal theory: Insights into long-range correlations of social and natural phenomena over the past two centuries. *Journal of the Royal Society Interface*; **9** (73): 1956–1964.
Gerlach M and Altmann EG (2013) Stochastic model for the vocabulary growth in natural languages. *Physical Review X*; **3** (2): 021006.
Google Inc. (2013) Ngram Viewer—Google Books. https://books.google.com/ngrams.
Gray RD and Atkinson QD (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*; **426** (6965): 435–439.
Greenfield PM (2013) The changing psychology of culture from 1800 through 2000. *Psychological Science*; **24** (9): 1722–1731.
Heaps HS (1978) *Information Retrieval. Computational and Theoretical Aspects*. Academic Press: New York.
Hughes JM, Foti NJ, Krakauer DC and Rockmore DN (2012) Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences*; **109** (20): 7682–7686.
Lieberman E, Michel J-B, Jackson J, Tang T and Nowak MA (2007) Quantifying the evolutionary dynamics of language. *Nature*; **449** (7163): 713–716.
Lin Y, Michel J-B, Aiden EL, Orwant J, Brockman W and Petrov S (2012) Syntactic annotations for the google books ngram corpus, in 'Proceedings of the ACL 2012 system demonstrations', Jeju, Republic of Korea, Association for Computational Linguistics, pp. 169–174.
Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, Pickett JP, Hoiberg D, Clancy D, Norvig P and Orwant J (2011) Quantitative analysis of culture using millions of digitized books. *Science*; **331** (6014): 176–182.
Montemurro MA (2001) Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Physica A*; **300** (3–4): 567–578.
Montemurro MA and Pury P (2002) Long-range fractals correlations in literary corpora. *Fractals*; **10**, 451–461.
Montemurro MA and Zanette DH (2011) Universal entropy of word ordering across linguistic families. *PLoS ONE*; **6** (5): e19875.
Morgan MS (1990) *The History of Econometric Ideas*. Cambridge University Press: New York.
Newman ME (2006) Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*; **103** (23): 8577–8582.
Nowak MA, Komarova NL and Niyogi P (2002) Computational and evolutionary aspects of language. *Nature*; **417** (6889): 611–617.
Pagel M (2009) Human language as a culturally transmitted replicator. *Nature Reviews Genetics*; **10** (6): 405–415.
Pagel M, Atkinson QD, Calude AS and Meade A (2013) Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences*; **110** (21): 8471–8476.
Pechenick EA, Danforth CM and Dodds PS (2015) Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE*; **10** (10): e0137041.
Petersen AM, Tenenbaum JN, Havlin S and Stanley HE (2012b) Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports*; **2**, 313.
Petersen AM, Tenenbaum JN, Havlin S, Stanley HE and Perc M (2012a) Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports*; **2**, 943.
Sigman M and Cecchi GA (2002) Global organization of the Wordnet lexicon. *Proceedings of the National Academy of Sciences*; **99** (3): 1742–1747.
Twenge JM, Campbell WK and Gentile B (2012) Increases in individualistic words and phrases in American books, 1960–2008. *PLoS ONE*; **7** (7): e40181.
Watts D and Strogatz S (1998) Collective dynamics of 'small-world' networks. *Nature*; **393** (393): 440–442.
Wolfram Research, Inc. (2016) *Mathematica, version 10.4*. Wolfram Research, Inc.: Champaign IL.
Zipf GK (1949) *Human Behavior and the Principle of Least Effort*. Addison-Wesley: Reading, MA.

Data availability

The datasets analysed during the current study are available in the Dataverse repository: http://dx.doi.org/10.7910/DVN/GMIXKB.

Additional information

Supplementary Information: accompanies this paper at http://www.palgrave-journals.com/palcomms

Competing interests: The Authors declare no competing financial interests.

Reprints and permission information is available at http://www.palgrave-journals.com/pal/authors/rights_and_permissions.html

How to cite this article: Montemurro MA and Zanette DH (2016) Coherent oscillations in word-use data from 1700 to 2008. *Palgrave Communications*. 2:16084 doi: 10.1057/palcomms.2016.84.