PLOS ONE

RESEARCH ARTICLE

Functional shortcuts in language co-occurrence networks

Woon Peng Goh[1,2], **Kang-Kwong Luke**[3], **Siew Ann Cheong**[2,4]*

1 Interdisciplinary Graduate School, Nanyang Technological University, Singapore, Singapore, **2** Complexity Institute, Nanyang Technological University, Singapore, Singapore, **3** School of Humanities, Nanyang Technological University, Singapore, Singapore, **4** School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, Singapore

* cheongsa@ntu.edu.sg

Citation: Goh WP, Luke K-K, Cheong SA (2018) Functional shortcuts in language co-occurrence networks. PLoS ONE 13(9): e0203025. https://doi.org/10.1371/journal.pone.0203025

Editor: Emilio Ferrara, University of Southern California, UNITED STATES

Received: November 16, 2016

Accepted: August 14, 2018

Published: September 11, 2018

Data Availability Statement: The data for this study are from the Uppsala Student English Corpus (USEC), the single-author corpus (SAC), and the Brown Corpus (BC). The Uppsala Student English Corpus (USEC) is a collection of essays collected by the Department of English, Uppsala University (see study in http://www.engelska.uu.se/research/english-language/electronic-resources/use/) and is freely hosted on http://ota.ox.ac.uk/desc/2457 under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. The single-author corpus (SAC) may be obtained by downloading The Complete Project Gutenberg Works of Jane Austen by Jane Austen from https://

Abstract

Human language contains regular syntactic structures and grammatical patterns that should be detectable in their co-occurence networks. However, most standard complex network measures can hardly differentiate between co-occurence networks built from an empirical corpus and a body of scrambled text. In this work, we employ a motif extraction procedure to show that empirical networks have much greater motif densities. We demonstrate that motifs function as efficient and effective shortcuts in language networks, potentially explaining why we are able to generate and decipher language expressions so rapidly. Finally we suggest a link between motifs and constructions in Construction Grammar as well as speculate on the mechanisms behind the emergence of constructions in the early stages of language acquisition.

Introduction

The pioneering linguist de Saussure defined language as "a system of interdependent terms in which the value of each term results solely from the simultaneous presence of the others" [1]. Under this definition, it becomes reasonable to employ the framework of complex networks in linguistic studies. The network approach is especially appropriate in analyzing the complex relationships among components in a complex system [2–4]. In the case of language, networks relate words or other linguistic components within the context of semantic, syntactic, co-occurring, or other types of relationships. Semantic networks have been shown to possess small-world and scale-free properties [5, 6]. The syntactic relationship between linguistic components has also been examined by building co-occurring word networks [7, 8] and with networks of syntactic dependencies [9, 10]. Such networks also display the same small-world and scale-free properties.

The works cited above have largely concentrated on the macro-structural properties of language networks. These global approaches, however, are unable to detect grammatical/syntactic structures in the networks. For instance, it has been found that measures like mean path length [11], mean degree [11, 12], and mean clustering coefficient [11, 12] are not significantly different in syntactic networks and non-syntactic networks derived from scrambled text. In fact,

www.gutenberg.org/ebooks/31100 under the license described in http://gutenberg.org/license. The Brown Corpus (BC) is accessed through the Python programming environment by installing the freely-available Natural Language Toolkit (NLTK, see http://www.nltk.org/ for installation instructions) package. Those interested would be able to access these data in the same manner as the authors by following the links provided. The authors did not have any special access privileges to the data that others would not have.

Funding: WPG thanks the Interdisciplinary Graduate School, Nanyang Technological University for the scholarship that supports his Ph. D. education. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Zipf's Law, the scale-free distribution of word frequencies said to be responsible for the scale-free topology of language networks, can be derived from various language models that do not consider syntax at all [13–15]. An exception is noted in [12] which introduced a selectivity measure that distinguished shuffled text from real ones. Given that the existence of grammatical patterns is universal in all human languages, should these patterns manifest in localized regularities in the language network as detectable micro-structures? If so, then the detection of such micro-structures becomes a generalized process to characterize the grammar of any language since the network approach is universal.

Ultimately, the macro and micro-structures of language exist to serve the purpose of efficient communication. Human languages are vastly more complex than any other forms of animal communication [16–19]. Tellingly, the vastness of the lexicon and the recursive application of structural patterns allow language an almost infinite creative potential in producing meaningful utterances [20, 21]. Despite the complexity and variety of language, the processes of creating and understanding expressions are accomplished remarkably rapidly [7]. This efficiency can be partially attributed to the small-world organization [22] in both semantic and syntactic networks [5, 7, 23–25]. If detectable micro-structures reside within language networks, is it possible that they also play a role accelerating the speed of network access and navigation?

In this work, we adopt an unsupervised learning algorithm developed by Solan *et al.* [26], called the Motif EXtraction (MEX) algorithm that identifies micro-structures called motifs from patterns in word-to-word networks of natural languages and other sequential data. The MEX algorithm has been shown to be capable of learning syntax of both artificial and natural languages. In the first instance, MEX extracted motifs from synthetic context-free grammar corpora that correspond very closely to their underlying production rules [26]. In natural language, the motifs learnt by MEX from the ATIS-2 English language corpus were used to generate novel sentences that were largely judged to be grammatical [26]. Here, we propose and demonstrate that motifs extracted by MEX in language networks are not only objects to represent syntax but also serve as effective and efficient navigational shortcuts in the production and deciphering of language. We also show that motif densities are drastically different between real corpora and their non-syntactic (i.e. scrambled) equivalents, examine how structural properties of motifs evolve through different embedding levels, and speculate on how these motifs arise during language acquisition.

Complex-network approaches to the study of languages

There is now a large body of work applying complex-network approaches to the study of languages. In this section, we will review how complex-network methods have been used to analyze the styles of different authors [27–29], analyze the multi-layer structure of languages [30–32], identify documents with similar contents [33–35], and in doing scientific research, what references should we consider and which are the most important ones [36].

In literary circles, it is common for an author to use a pen name instead of his or her real name. Some authors even use multiple pen names, depending on which genres, or which age groups of readers they are writing for. On the other hand, we can also have multiple for-hire authors, writing different books of a series under the same pen name. Naturally, we expect the habits of different authors to be different, and these would manifest themselves as different literary styles. To detect different writing styles, Segarra *et al.* constructed the network of *function words* (also called *stop words*) [27]. The links between function words are weighted, to distinguish different separations between the function words, and also to allow these separations to be averaged. Treating a weighted network as a Markov chain, and using the relative entropy as

the distance between Markov chains, the weighted networks between different texts can then be compared using hierarchical clustering. If there are not too many authors and styles, but many texts for each style, the different styles (and hence authors) can be accurately identified. In contrast, Amancio *et al*. removed stop words from a text to work only with semantically meaningful words, which they further mapped to their singular and infinitive forms before linking adjacent words to form a word-to-word network. They then combined network features with linguistic features such as word frequency, intermittency, *n*-grams, and used machine learning techniques to identify the authors of various texts with reasonable accuracy [28, 29].

Whether written or spoken, we can break words down into alphabets or phonemes, which are then aggregated into words, and in turn aggregated into phrases, sentences, and higher-level organizations where syntax and semantics emerge. Therefore, we should think of a language as multiple layers of linguistic entities interacting within layers and also between layers. In Ref. [30], Liu and Cong investigated the differences and similarities between the semantic, syntactic, co-occurrence, and phonemes layers of modern Chinese, while by comparing the multilayer structures of Croatian and English [31, 32], Martinčić-Ipšić and her co-workers found universal structural properties regardless of the language at the word-level layers, whereas at the syllabic subword-level, there are more language-dependent structural properties.

Going further, we might also wonder whether network-based approaches can help us get at the meaning behind a text. Traditional ways to do this would be to automatically identify keywords from a text, or to extract the list of *n*-grams from the text, so that to compare multiple texts at the semantic level, all we need to do is to compare the lists of keywords or the lists of *n*-grams. In fact, complex-network approaches can be used to discover the keywords [35]. To disambiguate between words with different meanings, Silva and Amancio combined traditional classifiers like those mentioned above, and pattern-based network classifiers, to demonstrate machine learning accuracies in excess of 70% [33]. Realizing that genre is frequently correlated with style, which manifests itself in the structure of the text, Amancio and co-workers used machine learning techniques to compute the semantic similarity between texts [34]. They found that topological measurements on the network in conjunction with semantic features give the best performance.

Finally, using the semantic comparison methods they have developed, Amancio *et al*. tested the idea of whether it is possible for researchers to do an automatic survey of the literature using a prescribed set of keywords to discover a corpus of related papers. By incorporating citation information for this set of papers, Amancio *et al*. were able to discover the subset of papers which can be considered seminal or highly related [36].

Syntactic network of languages

There are two general approaches in constructing the syntactic networks of languages. The first is with a syntactic dependency network [9, 10, 23] obtained usually from corpora that have been manually annotated with dependency trees [37]. The second is through means of a co-occurrence network [8, 11] which models the linear ordering of words (or other linguistic units) [38] in a corpus. We focus on the latter technique because the former requires expert guidance in generating the dependency trees whether it is done manually or automatically (algorithms for parsing require supervised training data). Using dependency networks also precludes the analysis of very large or new corpora, esoteric languages with no existing training sets, as well as languages with unknown grammars.

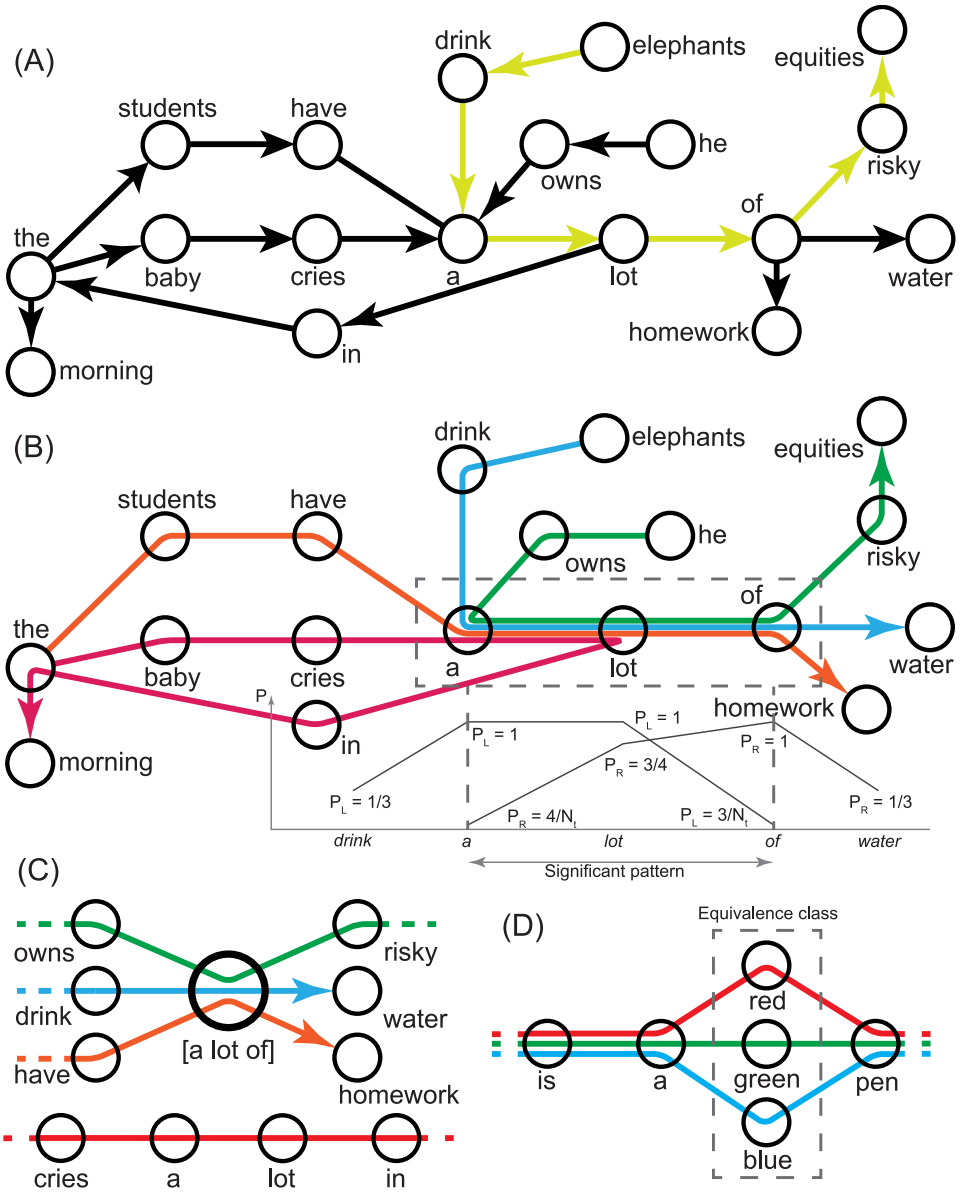

Fig 1. Toy co-occurrence network and pseudograph. (**A**) shows the co-occurrence network based on Sentences 1 to 4. We draw a directed edge between all adjacent word pairs. The unrestricted path from *elephants* to *equities* is highlighted in yellow. The unrestricted distance between the two words is 6. The pseudograph of the same toy corpus is shown in (**B**). Each sentence is represented by a different colored path on the network. A restricted path from *elephants* to *equities* does not exist. The box highlights the motif *a lot of*. It is marked by a fan-in of edges at the start (indicated by the sharp drop in the leftward extension probability P_L) and a fan-out at *of* (indicated by the sharp drop of the rightward extension probability P_R). A new node for the significant motif is created in (**C**) and the edges are routed through it. (**D**) identifies an equivalence class within a window $L_w = 4$. An equivalence class 'supernode' will be created and the edges are now routed through this new node. However, the unlike motif nodes, the new equivalence class node does not reduce distances between the terminal word nodes that lie on either side of it.

https://doi.org/10.1371/journal.pone.0203025.g001

The most straightforward method of constructing co-occurrence networks is to simply draw an edge (directed or undirected) for all pairs of neighbouring words (distance $d = 1$) in a corpus. With this, we risk missing out on interactions that take place at longer distances (i.e. $d \geq 2$). Fig 1A illustrates a co-occurrence network converted from a small toy corpus of the following 4 sentences:

1. Elephants drink a lot of water.
2. He owns a lot of risky equities.
3. The students have a lot of homework.
4. The baby cries a lot in the morning.

The simplicity of constructing co-occurrence networks means that it can be applied for any language. Even artificial ones can be cast on a network as long as the language is tokenizable. That is, the corpus can be split into tokens of linguistic units and expressed linearly with tokens arranged one after another. However, keeping only short-range correlations while discarding the long-range ones produces an over-generalized network. In the toy corpus, the words *elephants* and *equities* do not appear in the same sentence nor should they do so in normal English usage. Yet, in the network in Fig 1A, one can easily navigate along an unrestricted path p_{ur} from *elephants* to *equities* and produce a grammatical but unlikely phrase *elephants drink a lot of risky equities*. Clearly, not all possible paths in a co-occurrence network are accessed, or even allowed, in regular language use. In the later sections of this paper, mentions of language/linguistic networks will refer to such co-occurrence networks.

To overcome this over-generalization, Solan *et al*. [26] proposed a pseudograph which limits the number of allowable paths in the network. Instead of having edges that merely join pairs of words, edges in the pseudograph span multiple nodes (i.e. words). Multiple edges between nodes and loops are also allowed. In [26], each edge in the pseudograph connects all the words in a sentence in a linear order. Fig 1B shows the pseudograph of the same toy corpus. Since allowable paths must trace the available edges, a restricted path p_r (not to be confused with the rightward extension probability P_R) that leads from *elephants* to *equities* does not exist. Although the pseudograph does not lend itself well to traditional complex network analysis, the same work uses such a network to detect micro-structures called motifs. These motifs appear as coherent bundles of edges spanning short sequences of words and are book-ended at both ends with the convergence (fan-in) and divergence (fan-out) of edges. In the example in Fig 1B, the sub-sequence *a lot of* represents such a motif. The procedure of motif detection is described in detail in the methodology section. The motif is then embedded back into the corpus as a single linguistic unit (i.e. a single node, see Fig 1C), acting essentially as a shortcut by reducing network distances between words on opposite sides of the condensed phrase. Here, let us note that the hierarchy of motif embedding is related, but not entirely similar to the multi-layer network representation of language discussed earlier [30–32].

The idea of using network representations to model language goes back a long way (e.g. Jean Aitchison's book on the mental lexicon [39]). With recent advances in network modelling of human cognition, more and more is known about how word and syntactic structure are stored and accessed in the human mind [40]. Although there is little direct evidence of language networks, experiments have suggested that (i) degrees and local clustering coefficients in phonetic language networks [40] influence speed and accuracy in aural word recognition [41–43]; and that (ii) PageRank (a variant of network eigenvector centrality) of word nodes of the semantic network is better than raw frequencies in predicting human performance in language fluency tasks [44]. Moreover, it is thought that dependency crossings are minimized in the sequential order of sentences to optimize cognitive efficiency [45, 46]. The combined findings all indicate that networks can indeed be useful tools in modelling language processes in the human mind.

It has also been suggested that generating and deciphering language can be usefully modelled as navigation on language networks by means of various strategies [47] such as random

walks [48], switching random walks [49], random walks with memory [50], and using 'landmarks' with high network closeness centralities [51]. We further propose that linguistic networks contain shortcuts that optimizes their ease of navigation and access. The speed at which humans process language leads us to believe that syntactic networks are not navigated at a purely word-to-word level as this would entail longer path lengths ($\approx$ 20 words, the average length of a sentence) than can be stored in our much smaller working memories of 7 ± 2 objects [52, 53]. Shortcuts would allow us to traverse multiple nodes at a time, shortening the effective path lengths and utilizing less cognitive effort. We will demonstrate in this paper that the motifs detected are not just effective but also efficient network shortcuts.

Results

Motif detection in language networks

We applied the Motif EXtraction algorithm (MEX) [26] on three English corpora approximately 10^6 words in length. They are: i) the Uppsala Student English Corpus (USEC) (1007839 tokens, 22471 words, mean sentence length $\mu_{sent} = 19.6$) [54] of essays collected from 1999 to 2001 from students at Uppsala University in Sweden taking English as a second language which represents a learner's corpus; ii) the Brown Corpus (BC) (988331 tokens, 41018 words, $\mu_{sent} = 25.3$) [55] which is compiled from samples of mainly professionally written American English text complied in the 1960s to represent English usage at a high level of sophistication; and iii) a single-author corpus (SAC) (757542 tokens, 14456 words, $\mu_{sent} = 19.2$) of Jane Austen's writings freely available from Project Gutenberg [56] to represent also high proficiency but from a single user's perspective. The drop threshold parameter is set $\eta = 0.65$ and the context window length is $L_w = 6$ following [26] which selected these criteria based on the optimal trade-off between precision and recall.

MEX found the greatest number of patterns/motifs and equivalence classes (collectively termed objects) in the USEC. Fig 2A details the results. For example, at level 1, we detected $(6466 + 10191)/22471 = 0.74$ as many objects as there were terminal words in the original corpus (i.e. the motif density). It is followed closely by the SAC with 0.64 while the BC lags substantially with 0.16. We define corpus compression by measuring the change in the number of tokens ΔN_{toks} remaining in the corpus at each embedding level. It is greatest for the USEC which at level 5 contains $794922/1007839 = 0.79$ as many tokens as it did at level 0 (i.e. the normalized token count) and has a mean sentence length $\mu_{sent} = 15.4$. The SAC and the BC follow at 0.81 and 0.88 respectively. With other network measures (see Table 1), the BC also stands out with its very low clustering coefficient C [22] and density ρ when compared to the other two real corpora (1.1 vs. 3.1 and 4.4, all $\times 10^{-2}$).

We then tested MEX on artificially-generated corpora with little or no syntactic structures of real language. In Fig 2B, the word order in each sentence of the USEC was shuffled to produce the *shuffled corpus* (USEC-S) and the *POS-shuffled corpus* (USEC-PS) was generated by swapping words of the same Part-of-Speech (POS) category within the corpus. The POS categories for the USEC and SAC were created using a Maximum Entropy tagging algorithm implemented through Python's Natural Language Toolkit (NLTK) while the BC came with the POS tags included. As expected, only a negligible number of objects were identified in the scrambled corpora. The USEC-S and USEC-PS had motif densities of 0.01 and 0.05 respectively compared to 0.74 of the original USEC at level 1 and at level 2, the ratios of objects extracted from the USEC to its scrambled equivalents are even larger. However, we noted that the scrambled corpora have remarkably higher C, ρ, and average degree $\langle k \rangle$ (Table 1). The relatively low ρ and $\langle k \rangle$ in the USEC indicate selectivity in edge formation relative to scrambled corpora. Edges are selective in the sense that some connections in the unscrambled network

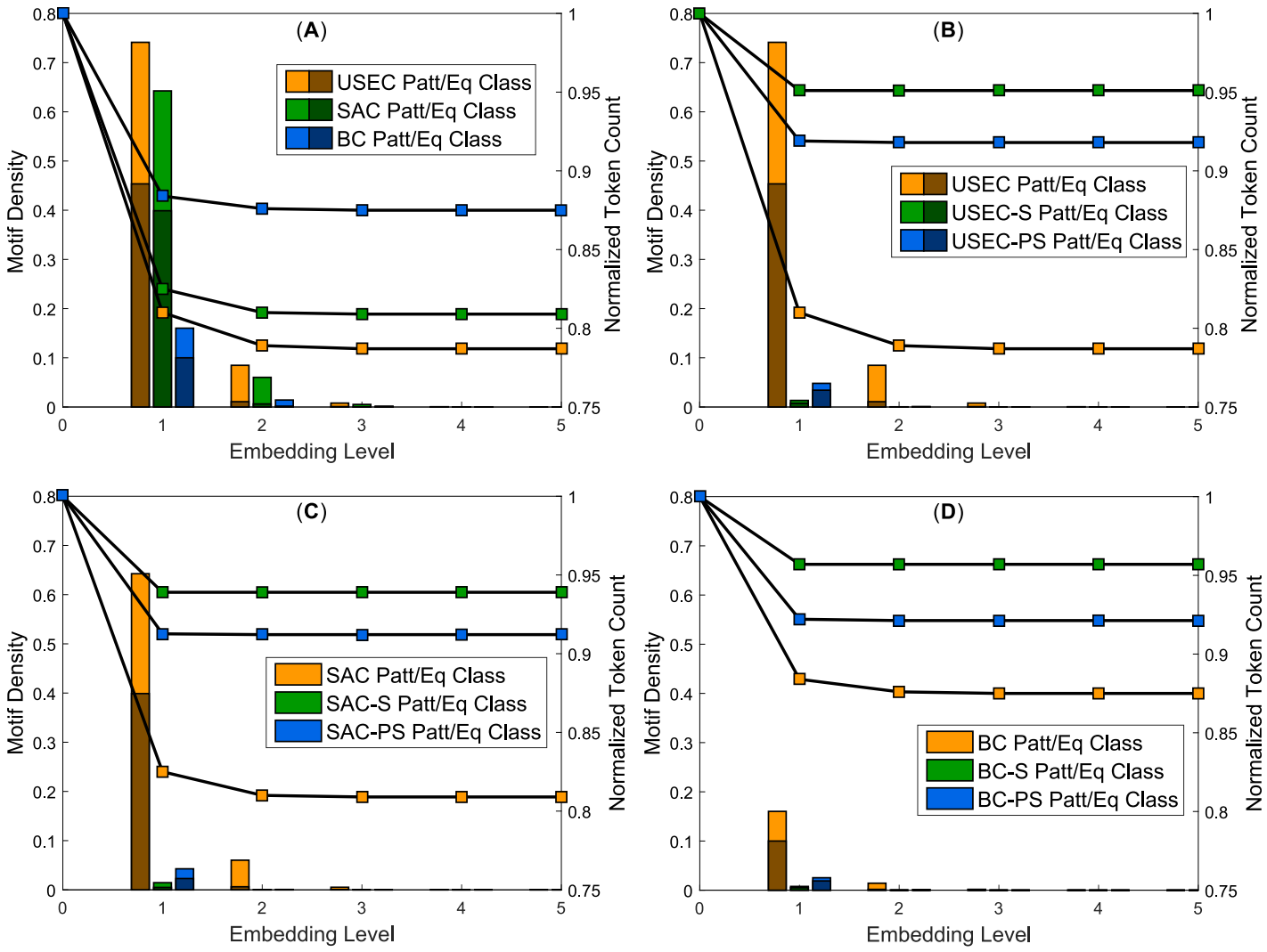

Fig 2. Results of MEX embedding. The bar graphs in the sub-plots show the motif densities (number of motifs divided by the number of original terminal words) and the line plots chart decrease in the number of tokens N^k_{toks}/N^0_{toks} as more motifs are embedded in the network. (**A**) gives the results for the real corpora (USEC, SAC, and BC) while (**B**), (**C**), and (**D**) show the difference between each real corpus, its shuffled equivalent and its POS-shuffled equivalent.

https://doi.org/10.1371/journal.pone.0203025.g002

(e.g. *to* to *the*) are in fact accessed many times in the corpus and after scrambling, these words become adjacent to a larger diversity of neighbours than they originally were.

The MEX algorithm identified motifs from overlaps in the word-to-word pseudograph. The BC is purposefully curated to include a diverse range of genres, subjects, and authorship to represent a wide range of American English usage and thus likely contains only a small number of overlaps in its network structure. Accordingly, the algorithm detected only a small number of motifs in this corpus. The USEC and the SAC, on the contrary, have much greater overlap densities. The USEC is built from students' essays written on a limited number of topics, with many essays per topic, guaranteeing content overlap. The SAC, composed of only 8 books from Jane Austen, contains both content and stylistic overlaps. It is apparent that selecting a corpus with the appropriate overlap density is more crucial for MEX to detect motifs than it is to choose a large corpus. The network measures C, ρ, $\langle k \rangle$, and assortativity r [57] proved to be unreliable in determining whether the word-to-word network contains MEX-detectable syntactic regularities.

Table 1. Table of network measures. The measurements are density ρ, average degree $\langle k \rangle$, clustering coefficient C, assortativity r, average minimum unrestricted distances $\langle min(d_{ur}) \rangle$ (i.e. distances along unrestricted paths as in Fig 1A), average minimum restricted distances $\langle min(d_r) \rangle$, and average mean restricted distances $\langle mean(d_r) \rangle$ (i.e. distances along restricted paths as in Fig 1B). For scrambled (appended with -S) and POS scrambled (appended with -PS) corpora, the values are given up to the precision not affected by fluctuations in the random scrambling.

Corpus	$\rho \times 10^4$	$\langle k \rangle$	$C \times 10^2$	r	$\langle min(d_{ur}) \rangle$	$\langle min(d_r) \rangle$	$\langle mean(d_r) \rangle$
USEC	5.1	23.0	3.05	−0.21	3.08	8.97	11.03
SAC	9.7	28.1	4.41	−0.26	2.89	16.10	20.70
BC	2.3	19.2	1.10	−0.19	3.09	9.79	11.48
USEC-S	8.4	37.5	7.66	−0.24	2.96	8.37	10.57
USEC-PS	7.8	34.9	5.43	−0.24	2.85	8.71	10.62
SAC-S	15.2	44.1	9.34	−0.29	2.83	15.52	20.24
SAC-PS	13.8	40.0	7.20	−0.29	2.78	15.73	20.32
BC-S	3.2	25.9	2.73	−0.18	3.03	9.44	11.18
BC-PS	2.9	23.7	1.61	−0.20	2.98	9.99	11.48

https://doi.org/10.1371/journal.pone.0203025.t001

In real corpora, increases in C, ρ, and $\langle k \rangle$ seem to correlate loosely with a larger motif density. However, the correlation is completely reversed with the scrambled corpora in the sense that they register hardly any objects in MEX despite possessing higher C, ρ, and $\langle k \rangle$. This agrees with the results of [11] which found that $\langle min(d_{ur}) \rangle$, $\langle k \rangle$, and C are slightly, but not significantly, higher for non-syntactic networks compared to their syntactic equivalents.

Network distances

Network distances will invariably shrink when a sequence of words is condensed into a single-node shortcut as it does under MEX. We quantify the evolution of network distances under MEX and compare it with an *equal-cost null model* (see methodology section) to demonstrate that MEX motifs are efficient in shrinking network distances. We consider distances only along paths p_r already existing on the pseudograph (i.e. the restricted distances). The reason is two-fold: i) as stated, many of the unrestricted paths are not accessed, or even allowed, in regular language use and ii) as shown in Table 1, unrestricted distances $\langle min(d_{ur}) \rangle$ are much smaller ($\approx 1/3$) than the observed restricted distances $\langle min(d_r) \rangle$ in the network. This is effectively what Margan *et al*. have found in Ref. [31].

The decrease of average mean network distances $\langle mean(d_r) \rangle$ are charted in Fig 3A and S2(A) and S2(B) Fig (see S2(C)–S2(E) Fig in for $\langle min(d_r) \rangle$). At level 5, the maximal layer of embedding, $\langle mean(d_r) \rangle$ for the USEC, SAC, and BC are respectively 8.571/11.027 = 0.777, 16.637/20.696 = 0.804, and 9.936/11.481 = 0.865 of the values of $\langle mean(d_r) \rangle$ at level 0 (i.e. the normalized distances). These fractional reductions seem large, but to convince ourselves that they are significant we need to compare them against the decrease of $\langle mean(d_r) \rangle$ from null models where the shortcuts are random, and thus meaningless. As we will explain the Methods section, for a proper comparison between empirical distances and null-model distances, something must be kept constant. Since it is extremely difficult to keep the numbers of new nodes and new edges constant, we constructed a family of null models where we can keep the *total cost* of adding new nodes and new edges the same as for the empirical network. Regardless of the node-formation cost Γ_{Node}, the distances of the MEX-embedded networks are always smaller than the null models (see Methods section). This is especially true when Γ_{Node} of the null models are set larger than 0.5. Null model distances shrink and approach the MEX results when $\Gamma_{Node} < 0.5$. For $\langle min(d_r) \rangle$, null models with $\Gamma_{Node} < 0.5$ sometimes even surpass MEX in distance reduction (e.g. in SAC and BC). It is not trivial to determine which Γ_{Node} yields the most realistic null model. In this, one may be guided by empirically calculating the ratio of

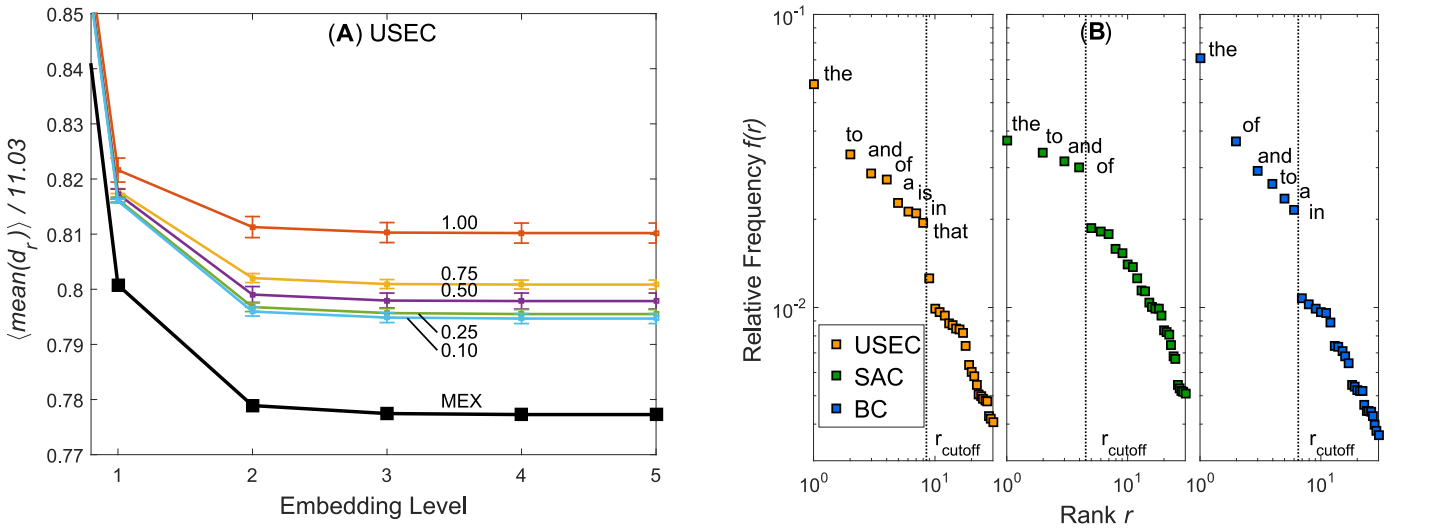

Fig 3. Shrinking distances in the USEC and ranked word frequencies. (**A**) show how $\langle mean(d_r)\rangle$ decreases in the USEC and how it compares to null models set at different cost parameters Γ_{Node}. (**B**) shows relative the word frequencies $f(r)$ against rank r and highlights the stop word cutoff in each corpus. The cutoffs mark sudden drops in the ranked word frequencies.

https://doi.org/10.1371/journal.pone.0203025.g003

new nodes to new edges created at each MEX level k and set $\Gamma_{Node}N(V_P^{(k)}) = \Gamma_{Edge}N(E_P^{(k)})$. This assumes that equal cost is expanded in node and edge creation. This yields Γ_{Node} between 0.82 to 0.97 which is well within the range where MEX shrinks the network more than the null model in both $\langle mean(d_r)\rangle$ and $\langle min(d_r)\rangle$. The choice of $\Gamma_{Node}/\Gamma_{Edge} > 0.5$ can also be motivated from neuropsychology—the creation of new memory that can be associated to previously acquired knowledge (i.e. edges) requires less effort than remembering a new and isolated piece of information (i.e. nodes) [58].

Taxonomy of MEX patterns

We also attempted to create a taxonomy of MEX-detected patterns. We use the POS tags attached to each token to create *POS templates* of MEX patterns. For example, the phrase *the united states* would have the template *DT (Determiner) JJ (Adjective) NN (Noun)* as would the phrase *a short man*. Significant POS templates are singled out by how their standard scores (z-score) measure against the null ensemble of patterns $V_P^{(k)*}$. The most significant POS templates at the first level is *DT (Determiner) NN (Noun) IN (Preposition)* (e.g. *a lot of*) with $z = 56.75$ and frequency $f = 371$. The next is *IN (Preposition) DT (Determiner)* (e.g. *in the*) (see Table 2 for results of the USEC and S1 and S2 Tables for the SAC and the BC). Additionally, the 6466 MEX pattern motifs at level 1 of the USEC can be represented by 1190 POS templates whereas the null model required twice the number of template types (2290). The relatively small set of POS templates used suggests that motifs at this level are in part characterized by a small class of grammatical/syntactic regularities. Significant templates are also detected in $G^{(2)}$ such as *[[IN DT] JJ NN]*. There are no significant ($z > 3$) POS templates beyond $G^{(2)}$.

Classification templates for patterns can also be constructed from *stop words* [59] which in computational linguistics refer to very common closed-class words such as prepositions, conjunctions, and pronouns with little semantic value. While the determination of POS tags demands expert input, *stop word templates* can be constructed from a purely data-driven approach. The list of stop words is not universal and varies among natural language processing

Table 2. Motif templates of the USEC. We present for levels 1 and 2 the top 3 ranked (in terms of Z-score) POS templates and stop word templates respectively. The Penn Treebank [60] POS tags are used here. Beyond level 2, there were no regularly used templates. For each level, we also gave the number of template types used by the motifs in the corpus and also the number of template types that appeared in randomly extracted motifs.

Rank	Template	Z	F	Example
Level 1 POS Templates: 6466 Motifs; 1190 Types; 2290 Null Types				
1	[DT NN IN]	56.75	371	[a lot of]
2	[IN DT]	41.70	434	[in the]
3	[DT VBZ]	34.00	81	[this is]
Level 2 POS Templates: 1662 Motifs; 1312 Types; 1442 Null Types				
1	[[IN DT] NN [IN DT]]	12.97	6	[[as a] {consequence, result} [of this]]
2	[[IN DT] JJ NN]	12.04	22	[[in the] first place]
3	[[IN DT] NN NN]	7.43	11	[[in the] right way]
Level 1 Stop Word Templates: 4038 Motifs; 83 Types; 98 Null Types				
1	_ the _ of _	39.77	227	[the age of]
2	_ a _ of _	26.54	110	[a lot of]
3	_ the _ to _	19.46	58	[the right to]
Level 2 Stop Word Templates: 353 Motifs; 24 Types; 38 Null Types				
1	_ of _ to _	7.88	3	[[in the] ages of {15, 16. . .} to]
2	_ of _	3.81	47	[[as a] matter of fact]
3	_ to _ to _	3.37	2	[to {quick, young} to]

https://doi.org/10.1371/journal.pone.0203025.t002

tools. We can, however, adopt an operational definition of stop words from empirical data. In Fig 3B, we plotted the ranked relative word frequencies $f(r)$ of each corpus on log-log plots and observed that there are sudden drops in frequency after $r = 8, 4, 6$ for respectively the USEC, SAC, and BC. This boundary marks a conservative rank cutoff r_{cutoff} to define an operational list of stop words for a particular corpus. It is entirely possible that stop words (in the closed-class definition sense) exist beyond these cutoffs but are not reflected in the trend of the data and thus cannot be unambiguously identified. For example, stop words like *the* and *to* are common for all three corpora and others like *is* is USEC-specific.

We employ these stop words to construct the *stop word templates*. Stop word templates are defined by replacing the non-stop words in a motif with a blank. For example, the stop word template of *the {reason, explanation} for* is _ *the* _ *for* _ with the inclusion of blanks also at the front and back. Of the 6466 MEX patterns at level 1 of the USEC, 4038 contain at least one stop word in them and can be classified by a template (Table 2). At $G^{(1)}$ of the USEC, we find significant templates in _ *the* _ *of* _ (e.g. *the age of*) ($z = 39.77, f = 227$) and _ *a* _ *of* _ (e.g. *a lot of*) ($z = 39.77, f = 227$). Similar templates are also found in the other two corpora (see S1 and S2 Tables). Beyond $G^{(1)}$, the significance and frequency of stop word templates fall drastically with very few registering $z > 3$. This suggests that at $G^{(>1)}$, motifs stop being organized around these operational lists of stop words. We also observe that at the first level $G^{(1)}$, stop words appear in MEX-detected motifs 58 to 83% more frequently than by chance (see S3 Table). The trend is inverted at higher levels $G^{(>1)}$ as the rates of stop word usage decrease below that of random noise (e.g. at level 2, 16 to 30% below null rate). Although, stop word templates cannot classify patterns as precisely as POS templates (fewer possible template types), they are nevertheless a good unsupervised alternative.

Discussion

In linguistics, the theories of Construction Grammar (CxG) describe a family of grammar models in which constructions are the basic units of grammar [61–63]. Constructions, in this

framework, are *pairings of form and function* [62, 63]. Form refers to the morphological and syntactic features of the construction while function specifies its semantic features [64]. Consider the idiom *kick the bucket* [62]. One cannot derive its meaning by considering its syntax (the 'X does something to Y' form) and its component words separately. In CxG, not only idiomatic expressions, but all constructions are said to be form-function pairings. There is also "a continuum from schematic complex constructions to substantive atomic constructions" [65]. Isolated words and complex grammatical arrangements are treated on equal footing in CxG.

One of the processes through which children acquire constructions is entrenchment [66]. Entrenchment is when a person performs a task successfully enough times, the way they perform this task becomes habitual and automatic. Similarly, MEX detect motifs in language pseudographs by searching for repeated sub-sequences. They are reminiscent of constructions in CxG. Take, for example, the generalized pattern *[i {feel, think, believe} that]* found in the USEC. All three variants of the pattern are identical in their functions of expressing an opinion. Even though it shares an identical form with a pattern like *[i suggest that]*, MEX considers them to be different entities due to their usage in different contexts. Such form-function pairings in MEX motifs are features found also in CxG. Of course, it would be presumptuous to claim that all motifs in MEX are constructions, or that all constructions in the corpus are detectable with MEX. Nevertheless, MEX remains a useful tool to quickly and crudely process a corpus for possible constructions.

Frequently used constructions become entrenched in the mental grammar [64]. These constructions become mental routines directly accessed without invoking higher schemas [64]. This is consistent with our earlier claim that the language network has to contain shortcuts to account for how rapidly we navigate it. In MEX, significant motifs are similarly entrenched in the network as condensed routines. As shown, motifs reduce observed network distances d_r effectively and efficiently. We believe that the creation of such shortcuts in the mental lexicon comes at the expense of using more memory to store the new linguistic units and the new associations that come with it. This cost, modelled in Γ, is offset by the reduction of network distances which we associate with an increase in language processing speed. Earlier, we alluded to the usefulness of modelling language processes in the human mind as network processes. Here, we suggest that these networks are not merely a simple matter of concatenating adjacent words to form strings, but rather contain complex shortcut structures that can be deduced from motif detection.

We can also speculate on the mechanisms that guide the formation of constructions by investigating the properties of MEX motifs. We showed that incidence rates of stop words in level-1 motifs are higher than expected from the frequencies of the stop words themselves. In syntactic networks of natural languages, stop words are hubs owing to their high usage frequency and their flexible combinatorial potential [24]. It has been shown that stop words only become hubs in the syntax network of a child after the *syntactic spurt* [67] stage of language acquisition [23]. If we assume that level-1 motifs correspond to constructions acquired in early stages of language development, this study lends credence to our observation that stop words are important building blocks of low-level motifs. Therefore, depending on the questions we are interested in, we should not always remove stop words from our word co-occurrence network (as is done, for example, in Refs. [28, 29]).

In linguistics, stop words fall under the category of *function words* [68] which can be prepositions, pronouns, auxiliary verbs, conjunctions, articles, or particles. Function words by themselves have little lexical meaning and, instead, serve to mediate and/or to emphasize the interaction between words. We believe that function words arise necessarily out of cognitive limitations. Before they undergo the syntactic spurt, toddlers have small one-word lexicons [67] which can be easily searched through when producing or deciphering utterances. As the

lexicons grow, and as the children start to produce 2-to-3 word phrases [67], the numbers of possible expressions explode combinatorially. The same linear search strategy becomes too slow for real-time linguistic interactions. Function words are constrained by the word category preceding them and they in turn constrain the word category following them. For example, a toddler may say 'put table/bed/box' to mean 'put the toy on the table/bed' or 'in the box'. With the introduction of a function word like *on*, after saying 'put on', it is much more probable to produce *table* or *bed* than *box*. As such, the function word *on* reduces the search space of the possible productions after it.

Methods

Motif EXtraction algorithm

The MEX (Motif EXtraction) algorithm was developed by Solan *et al.* [26] to automatically extract patterns and syntax from language corpora. Consider a corpus of N_{sents} sentences of varying lengths totaling N_{toks} tokens made up of N_{words} unique words. It is useful to visualize this corpus as a pseudograph where each word is represented by a node and each sentence is a directed edge or path going through multiple nodes. As described, motifs are coherent sub-sequences of nodes where a number of edges bundle up and are bookended with fan-ins and fan-outs of edges.

We define a path in this graph as $(e_i; e_j) = (e_{i+1}, e_{i+2},\ldots,e_j)$ where each e_k represents some word node w_l. The *extension probabilities* of such a path are

$$P_R\left(e_i; e_j\right) = \frac{\ell(e_i; e_j)}{\ell(e_i; e_{j-1})}, j > i \tag{1}$$

and

$$P_L\left(e_j; e_i\right) = \frac{\ell(e_i; e_j)}{\ell(e_{i+1}; e_j)}, j > i. \tag{2}$$

The function $\ell(e_i; e_j)$ returns the number of edges that traverse the sequence $(e_i, e_{i+1},\ldots,e_j)$. Therefore, the equation for $P_R(e_i;e_j)$ describes the probability that a sub-path $(e_i, e_{i+1},\ldots,e_{j-1})$ is found to extend rightward (or forward) to word e_j. Similarly, $P_L(e_j;e_i)$ measures the leftward (or backward) extension probability i.e. the probability that a sub-path $(e_{i+1}, e_{i+2},\ldots,e_j)$ is preceded by word e_i. $P_R(e_i;e_i) = P_L(e_i;e_i)$ is simply the occurrence probability of e_i. As illustrated in Fig 1B, the start and end of a significant motif are marked by sudden drops in their rightward and leftward extension probabilities. There also is precedence of using such extension probabilities to analyze transitions in phonological linguistic elements in [69, 70]. The *drop ratios* are defined as

$$D_R\left(e_i; e_j\right) = \frac{P_R(e_i; e_j)}{P_R(e_i; e_{j-1})}, j > i \tag{3}$$

and

$$D_L\left(e_j; e_i\right) = \frac{P_L(e_j; e_i)}{P_L(e_j; e_{i+1})}, j > i. \tag{4}$$

When $D < \eta$, the drop threshold cutoff, we consider the sub-sequence a significant motif. A new pattern node $e_{patt} = [e_{i+1}, e_{i+2},\ldots,e_j]$ which contains the condensed sequence is created. Paths that previously went through $(e_{i+1}, e_{i+2},\ldots,e_j)$ are now routed through the e_{patt} pattern 'supernode' like in Fig 1C. When there are overlapping significant sub-sequences, we prioritize

the one with greater significance based on the smallest combined cumulative binomial probability $B_{comb} = B_R + B_L$ where

$$B_R(e_i; e_j) = \sum_{x=0}^{\ell(e_i;e_j)} Binom(\ell(e_i; e_{j-1}), x, \eta P(e_i; e_{j-1})), j > i \quad (5)$$

and

$$B_L(e_j; e_i) = \sum_{x=0}^{\ell(e_j;e_i)} Binom(\ell(e_{i+1}; e_j), x, \eta P(e_{i+1}; e_j)), j > i \quad (6)$$

where

$$Binom(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}. \quad (7)$$

Intuitively, low individual probabilities B_R and B_L mean that there are *fewer* extended subsequences than expected and thus the drop ratios are more significant. MEX can also accommodate *equivalence classes* of linguistic units that share the same context. Fig 1D shows an example of an equivalence class contained within a context window of length $L_w = 4$. The slot for the equivalence class can exist at any position of a context window except at the edges. The equivalence class is then merged into an equivalence class 'supernode'. As seen, the equivalence class node does not shrink distances in the network.

In our implementation of MEX, we consider the network formed from the original corpus of terminal words to be the 0-th level graph with its collection of vertices and edges ($G^{(0)} = (V^{(0)}, E^{(0)})$). From this, we scan the graph first for all possible level 1 equivalence classes (i.e. the generalization step) and embed them to $G^{(0)}$, obtaining $G^{(1)-OnlyEC}$. We then seek for pattern candidates from all possible sub-sequences in $G^{(1)-OnlyEC}$ and rank the candidates by B_{comb}. The patterns are embedded into the network starting from the most significant and the process is repeated until no more level-1 pattern vertices $V_P^{(1)}$ are found (i.e. the embedding step). This culminates in the level 1 network $G^{(1)}$. This ordering of steps enables the equivalence class detection and pattern detection procedures to bootstrap each other to create more equivalence classes and pattern motifs. The algorithm then moves on to construct $G^{(2)}$ and so on until no more new objects are identified. This recursive search creates a kind of hierarchy in the embedding of motifs in the network where motifs at a higher embedding layer are compositions of lower-level motifs akin to hierarchies in syntactic structures [71]. For instance, short sequences of words form simple noun-phrase or verb-phrase type motifs which are assembled into complex sentence segments as a higher-level motif.

Equal-cost null model for network shortcuts

As we perform the MEX procedure on a linguistic network, we create motif 'supernodes' that act as distance-reducing bridges (see Fig 1C where, for example, the distance from *drink* to *water* was reduced from 4 to 2 with the creation of the *[a lot of]* shortcut). We believe that there is a 'cost' involved in forming such shortcuts. To calculate this cost, recall that when we derive the pseudograph $G^{(k)}$ from $G^{(k-1)}$ using MEX, $N(V_P^{(k)})$ number of motifs are found and embedded into the $G^{(k)}$ pseudograph as shortcut 'supernodes'. Concurrently, $N(E_P^{(k)})$ edges are also created to link these new nodes to existing ones. We hypothesize that these new network objects are created in the mental lexicon at the cost of

$$\Gamma(G^{(k)}, G^{(k-1)}) = \Gamma_{Node} N(V_P^{(k)}) + (1 - \Gamma_{Node}) N(E_P^{(k)}). \quad (8)$$

where Γ_{Node} and $\Gamma_{Edge} = 1-\Gamma_{Node}$, both in [0, 1], model the creation costs per node and per edge respectively. $\Gamma(G^{(k)}, G^{(k-1)})$ thus gives the total embedding cost when MEX is performed on the $G^{(k-1)}$ network.

To determine if such MEX-detected shortcuts are in fact *efficient* in shrinking distances, we compare them against a null model where shortcuts are randomly created. We start with $G^{(0)\dagger} = G^{(0)}$, which is an instance of the null model of $G^{(0)}$. To obtain $G^{(1)\dagger}$, we keep selecting random sub-sequences in the network to be condensed into shortcuts until $\Gamma(G^{(k)\dagger}, G^{(k-1)\dagger}) = \Gamma(G^{(k)}, G^{(k-1)})$ i.e. the total cost of creating the random shortcuts becomes equal to that of the MEX shortcuts. The subsequent levels $G^{(2)\dagger}$, $G^{(3)\dagger}$, etc. are similarly derived from $G^{(1)\dagger}$, $G^{(2)\dagger}$, etc.

Different $G^{(k)\dagger}$ null models can be obtained by setting different values of Γ_{Node}. When the ratio $\Gamma_{Node}/(1-\Gamma_{Node})$ is high ($\Gamma_{Node} \approx 1$ and $\Gamma_{Edge} \approx 0$), the null model $G^{(k)\dagger}$ will contain a similar number of new shortcut nodes to the MEX-processed corpus (i.e. $N(V_P^{(k)\dagger}) \approx N(V_P^{(k)})$). In this scenario, $G^{(k)\dagger}$ has somewhat fewer new edges (i.e. $N(E_P^{(k)\dagger}) < N(E_P^{(k)})$) since the MEX criteria implies that each new shortcut sub-sequence is traversed by more than one sentence path (i.e. more than one edge incident on the new node) whereas the randomly created shortcuts are often only used by a single sentence. Conversely, a low $\Gamma_{Node}/(1-\Gamma_{Node})$ ratio results in a $G^{(k)\dagger}$ model where the total number of edges incident on the randomly created shortcuts is similar to the original ($N(E_P^{(k)\dagger}) \approx N(E_P^{(k)})$) and, by the same argument, a greater number of new nodes (i.e. $N(V_P^{(k)\dagger}) > N(V_P^{(k)})$).

Null rate of pattern properties

In examining the properties of the pattern motifs detected by MEX, it is essential to compare these quantities to a null rate i.e. the properties of the ensemble of patterns detected in level k if, instead of being detected with the MEX criteria, they were randomly defined. We take $G^{(k)-OnlyEC}$, which is network $G^{(k)}$ after embedding the equivalence class nodes but *before* patterns are embedded using MEX, and define the ensemble of null model patterns $V_P^{(k)*}$ by selecting an equal-length random sub-sequence in this network for every actual motif subsequence found by MEX. The null values of the pattern properties are then computed from this ensemble.

Supporting information

S1 Fig. Embedding example. An example sequence taken from the USEC is shown here at different levels of embedding. At $G^{(0)}$, only terminal word nodes exist. At $G^{(0)-OnlyEC}$, some of the terminal word nodes are embedded inside equivalence class supernodes such as *totally* being embedded within {*quite, totally, entirely*}. At $G^{(1)}$, the nodes merged into 3 separate level-1 pattern motifs and at $G^{(2)}$ these 3 motifs combine to form a level-2 pattern motif.
(PDF)

S2 Fig. Decrease of network distances under motif embedding. A and **B** shows the decrease of $\langle mean(d_r) \rangle$ in the SAC and BC and how they compare to null models set at different cost parameters. Sub plots (**C**-**E**) chart the decrease of $\langle min(d_r) \rangle$ and their null models for the USEC, SAC, and BC respectively.
(PDF)

S1 Table. POS templates of motifs in the SAC. This table is interpreted in the same manner as Table 2.
(PDF)

S2 Table. POS templates of motifs in the BC. The BC uses a different list of POS tags which can be obtained from http://www.comp.leeds.ac.uk/ccalas/tagsets/brown.html. This table is interpreted in the same manner as Table 2.
(PDF)

S3 Table. Stop word templates of motifs in the SAC.
(PDF)

S4 Table. Stop word templates of motifs in the BC.
(PDF)

S5 Table. Table of pattern properties. Here we show the properties of the patterns extracted by MEX for the first 3 levels. $\langle N_{len(P)=2} \rangle$ gives the mean occurrence frequency of length-2 patterns and $\langle N_{len(P)>2} \rangle$ is for patterns with lengths greater than 2. F_{SWinP}, F_{CinP}, and F_{PinP} are the proportions of objects in the patterns that are stop words, classes, and lower-level patterns respectively. The values are presented with the difference between the observed values and the null values in parentheses together with the error margin. For example, $\langle N_{len(P)=2} \rangle$ for the USEC at level 1 is 44.3 and the null model yields 44.3 − 1.3 = 43.0 with and error margin of ±0.3.
(PDF)

S6 Table. Table of stop words. Operationally-defined stop words for each of the 3 corpora is given here together with the the part(s) of speech they belong to.
(PDF)

Acknowledgments

WPG thanks the Interdisciplinary Graduate School, Nanyang Technological University for the scholarship that supports his Ph.D. education. The authors thank our colleague J. Stephen Lansing for his suggestion that linguistic motifs might exist in a hierarchy and his helpful comments in the preparation of the manuscript.

Author Contributions

Conceptualization: Woon Peng Goh, Kang-Kwong Luke.

Formal analysis: Woon Peng Goh.

Methodology: Woon Peng Goh.

Supervision: Siew Ann Cheong.

Writing – original draft: Woon Peng Goh, Kang-Kwong Luke, Siew Ann Cheong.

Writing – review & editing: Woon Peng Goh, Kang-Kwong Luke, Siew Ann Cheong.

References

1. De Saussure F, Baskin W. Course in general linguistics. Columbia University Press; 2011.
2. Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature. 2005; 435(7043):814–818. https://doi.org/10.1038/nature03607 PMID: 15944704
3. Bullmore E, Sporns O. Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience. 2009; 10(3):186–198. https://doi.org/10.1038/nrn2575 PMID: 19190637

4. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science. 2002; 298(5594):824–827. https://doi.org/10.1126/science.298.5594.824 PMID: 12399590
5. Sigman M, Cecchi GA. Global organization of the Wordnet lexicon. Proceedings of the National Academy of Sciences. 2002; 99(3):1742–1747. https://doi.org/10.1073/pnas.022341799
6. Motter AE, de Moura AP, Lai YC, Dasgupta P. Topology of the conceptual network of language. Physical Review E. 2002; 65(6):065102.
7. Ferrer i Cancho R, Solé RV. The small world of human language. Proceedings of the Royal Society of London B: Biological Sciences. 2001; 268(1482):2261–2265. https://doi.org/10.1098/rspb.2001.1800
8. Masucci A, Rodgers G. Network properties of written human language. Physical Review E. 2006; 74 (2):026102. https://doi.org/10.1103/PhysRevE.74.026102
9. Ferrer i Cancho R, Solé RV, Köhler R. Patterns in syntactic dependency networks. Physical Review E. 2004; 69(5):051915. https://doi.org/10.1103/PhysRevE.69.051915
10. Liu H. The complexity of Chinese syntactic dependency networks. Physica A: Statistical Mechanics and its Applications. 2008; 387(12):3048–3058. https://doi.org/10.1016/j.physa.2008.01.069
11. Liu H, Hu F. What role does syntax play in a language network? EPL (Europhysics Letters). 2008; 83 (1):18002. https://doi.org/10.1209/0295-5075/83/18002
12. Masucci A, Rodgers G. Differences between normal and shuffled texts: structural properties of weighted networks. Advances in Complex Systems. 2009; 12(01):113–129. https://doi.org/10.1142/S0219525909002039
13. Dorogovtsev SN, Mendes JFF. Language as an evolving word web. Proceedings of the Royal Society of London B: Biological Sciences. 2001; 268(1485):2603–2606. https://doi.org/10.1098/rspb.2001.1824
14. Ferrer i Cancho R, Solé RV. Least effort and the origins of scaling in human language. Proceedings of the National Academy of Sciences. 2003; 100(3):788–791. https://doi.org/10.1073/pnas.0335980100
15. Corominas-Murtra B, Hanel R, Thurner S. Understanding scaling through history-dependent processes with collapsing sample space. Proceedings of the National Academy of Sciences. 2015; 112(17):5348–5353. https://doi.org/10.1073/pnas.1420946112
16. Solé RV. Phase Transitions. Primers in Complex Systems. Princeton University Press; 2011. Available from: https://books.google.com.pe/books?id=YcEY1OwDgMkC.
17. Maynard Smith J, Szathmary E. The major evolutionary transitions. Nature. 1995; 374:227–232. https://doi.org/10.1038/374227a0
18. Bickerton D. Language and human behavior. University of Washington Press; 1995.
19. Hockett CF. The origin of speech. Scientific American. 1960; 203:88–96. https://doi.org/10.1038/scientificamerican0960-88
20. Chomsky N. Language and mind. Cambridge University Press; 2006.
21. Hauser MD. The evolution of communication. MIT press; 1996.
22. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998; 393(6684):440–442. https://doi.org/10.1038/30918 PMID: 9623998
23. Corominas-Murtra B, Valverde S, Solé RV. The ontogeny of scale-free syntax networks: phase transitions in early language acquisition. Advances in Complex Systems. 2009; 12(03):371–392. https://doi.org/10.1142/S0219525909002192
24. Solé RV, Corominas-Murtra B, Valverde S, Steels L. Language networks: Their structure, function, and evolution. Complexity. 2010; 15(6):20–26.
25. Borge-Holthoefer J, Arenas A. Semantic networks: structure and dynamics. Entropy. 2010; 12 (5):1264–1302. https://doi.org/10.3390/e12051264
26. Solan Z, Horn D, Ruppin E, Edelman S. Unsupervised learning of natural languages. Proceedings of the National Academy of Sciences. 2005; 102(33):11629–11634. https://doi.org/10.1073/pnas.0409746102
27. Segarra S, Eisen M, Ribeiro A. Authorship attribution through function word adjacency networks. IEEE Transactions on Signal Processing. 2015; 63(20):5464–5478. https://doi.org/10.1109/TSP.2015.2451111
28. Amancio DR. Authorship recognition via fluctuation analysis of network topology and word intermittency. Journal of Statistical Mechanics: Theory and Experiment. 2015; 2015(3):P03005. https://doi.org/10.1088/1742-5468/2015/03/P03005
29. Amancio DR. A complex network approach to stylometry. PloS one. 2015; 10(8):e0136076. https://doi.org/10.1371/journal.pone.0136076 PMID: 26313921

30. Liu HCJ. Empirical characterization of modern Chinese as a multi-level system from the complex network approach. Journal of Chinese Linguistics. 2014;p. 1–38.
31. Margan D, Martinčić-Ipšić S, Meštrović A. Network differences between normal and shuffled texts: Case of Croatian. In: Complex Networks V. Springer; 2014. p. 275–283.
32. Margan D, Martinčić-Ipšić S, Meštrović A. Multilayer network of language: A unified framework for structural analysis of linguistic subsystems. Physica A: Statistical Mechanics and its Applications. 2016; 457:117–128. https://doi.org/10.1016/j.physa.2016.03.082
33. Silva TC, Amancio DR. Word sense disambiguation via high order of learning in complex networks. EPL (Europhysics Letters). 2012; 98(5):58001. https://doi.org/10.1209/0295-5075/98/58001
34. Amancio DR, Oliveira ON Jr, Costa LdF. Structure—semantics interplay in complex networks and its effects on the predictability of similarity in texts. Physica A: Statistical Mechanics and its Applications. 2012; 391(18):4406–4419. https://doi.org/10.1016/j.physa.2012.04.011
35. Beliga S, Meštrović A, Martinčić-Ipšić S. An overview of graph-based keyword extraction methods and approaches. Journal of information and organizational sciences. 2015; 39(1):1–20.
36. Amancio DR, Nunes MdGV, Oliveira O, Costa LdF. Using complex networks concepts to assess approaches for citations in scientific papers. Scientometrics. 2012; 91(3):827–842. https://doi.org/10.1007/s11192-012-0630-z
37. Hudson R. An introduction to word grammar. Cambridge University Press; 2010.
38. Tesnière L. Eléments de syntaxe structurale. Klincksieck, Paris; 1959.
39. Aitchison J. Linguistics. Teach yourself books. Hodder and Stoughton; 1978. Available from: https://books.google.com.sg/books?id=J4Z-QgAACAAJ.
40. Beckage NM, Colunga E. Language networks as models of cognition: Understanding cognition through language. In Towards a Theoretical Framework for Analyzing Complex Linguistic Networks. 2015;p. 3–28.
41. Luce PA, Pisoni DB. Recognizing spoken words: The neighborhood activation model. Ear and hearing. 1998; 19(1):1. https://doi.org/10.1097/00003446-199802000-00001 PMID: 9504270
42. Chan KY, Vitevitch MS. The influence of the phonological neighborhood clustering coefficient on spoken word recognition. Journal of Experimental Psychology: Human Perception and Performance. 2009; 35(6):1934. https://doi.org/10.1037/a0016902 PMID: 19968444
43. Vitevitch MS, Ercal G, Adagarla B. Simulating retrieval from a highly clustered network: Implications for spoken word recognition. Frontiers in psychology. 2011; 2:369. https://doi.org/10.3389/fpsyg.2011.00369 PMID: 22174705
44. Griffiths TL, Steyvers M, Firl A. Google and the mind: Predicting fluency with PageRank. Psychological Science. 2007; 18(12):1069–1076. https://doi.org/10.1111/j.1467-9280.2007.02027.x PMID: 18031414
45. i Cancho RF. Why do syntactic links not cross? EPL (Europhysics Letters). 2006; 76(6):1228. https://doi.org/10.1209/epl/i2006-10406-0
46. Liu H. Dependency distance as a metric of language comprehension difficulty. Journal of Cognitive Science. 2008; 9(2):159–191. https://doi.org/10.17791/jcs.2008.9.2.159
47. Baronchelli A, Ferrer-i Cancho R, Pastor-Satorras R, Chater N, Christiansen MH. Networks in cognitive science. Trends in cognitive sciences. 2013; 17(7):348–360. https://doi.org/10.1016/j.tics.2013.04.010 PMID: 23726319
48. Borge-Holthoefer J, Arenas A. Categorizing words through semantic memory navigation. The European Physical Journal B. 2010; 74(2):265–270. Available from: http://dx.doi.org/10.1140/epjb/e2010-00058-9.
49. Goñi J, Martincorena I, Corominas-Murtra B, Arrondo G, Ardanza-Trevijano S, Villoslada P. Switcher-random-walks: A cognitive-inspired mechanism for network exploration. International Journal of Bifurcation and Chaos. 2010; 20(03):913–922. https://doi.org/10.1142/S0218127410026204
50. Capitán JA, Borge-Holthoefer J, Gómez S, Martinez-Romo J, Araujo L, Cuesta JA, et al. Local-based semantic navigation on a networked representation of information. PloS one. 2012; 7(8):e43694. https://doi.org/10.1371/journal.pone.0043694 PMID: 22937081
51. Sudarshan Iyengar S, Veni Madhavan C, Zweig KA, Natarajan A. Understanding human navigation using network analysis. Topics in cognitive science. 2012; 4(1):121–134. https://doi.org/10.1111/j.1756-8765.2011.01178.x PMID: 22253185
52. Miller GA. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review. 1956; 63(2):81–97. https://doi.org/10.1037/h0043158 PMID: 13310704
53. Christiansen MH, Chater N. The now-or-never bottleneck: A fundamental constraint on language. Behavioral and Brain Sciences. 2016; 39:e62. https://doi.org/10.1017/S0140525X1500031X PMID: 25869618

54. Axelsson MW. USE—the Uppsala Student English corpus: an instrument for needs analysis. ICAME Journal. 2000; 24:155–157.
55. Francis WN. A standard corpus of edited present-day American English. College English. 1965; 26 (4):267–273. https://doi.org/10.2307/373638
56. Hart M. Project gutenberg. Project Gutenberg; 1971.
57. Newman ME. Mixing patterns in networks. Physical Review E. 2003; 67(2):026126. https://doi.org/10.1103/PhysRevE.67.026126
58. Anderson JR, Bower GH. Human associative memory. Psychology Press; 2014.
59. Lawrence S, Giles CL. Searching the world wide web. Science. 1998; 280(5360):98–100. https://doi.org/10.1126/science.280.5360.98 PMID: 9525866
60. Marcus MP, Marcinkiewicz MA, Santorini B. Building a large annotated corpus of English: The Penn Treebank. Computational linguistics. 1993; 19(2):313–330.
61. Goldberg AE. Constructions: a new theoretical approach to language. Trends in cognitive sciences. 2003; 7(5):219–224. https://doi.org/10.1016/S1364-6613(03)00080-9 PMID: 12757824
62. Fillmore CJ. The mechanisms of a construction grammar. In: Annual Meeting of the Berkeley Linguistics Society. vol. 14; 1988. p. 35–55.
63. Croft W. Radical construction grammar: Syntactic theory in typological perspective. Oxford University Press; 2001.
64. Diessel H. The acquisition of complex sentences. Cambridge University Press; 2004.
65. Croft W. Construction grammar. The Oxford handbook of cognitive linguistics. 2007;p. 463–508.
66. Tomasello M. Acquiring linguistic constructions. Handbook of child psychology. 2006;.
67. Radford A. Syntactic theory and the acquisition of English syntax: The nature of early child grammars of English. Blackwell; 1990.
68. Fries CC. The structure of English: An introduction to the construction of English sentences. Longman; 1973.
69. Hockett CF. A manual of phonology. 11. Waverly Press; 1955.
70. Wang WY. The measurement of functional load. Phonetica. 1967; 16(1):36–54. https://doi.org/10.1159/000258556
71. De Beule J, Steels L. Hierarchy in fluid construction grammars. In: Annual Conference on Artificial Intelligence. Springer; 2005. p. 1–15.