An Empirical Study of Language Syllabification using Syllabary and Lexical Networks

Rusali Saha
Texas A&M University
rs0921@tamu.edu

Yannick Marchand
Faculty of Computer Science
Department of Pyschology and Neuroscience
Dalhousie University
marchand@cs.dal.ca

Abstract

Language syllabification is the separation of a word into written or spoken syllables. The study of syllabification plays a pivotal role in morphology and there have been previous attempts to study this phenomenon using graphs or networks. Previous approaches have claimed through visual estimation that the degree distribution of language networks follows the Power Law distribution, however, there have not been any empirically grounded metrics to determine the same. In our study, we implement two kinds of language networks, namely, syllabary and lexical networks, and investigate the syllabification of four European languages: English, French, German and Spanish using network analysis and examine their small-world, random and scale-free nature. We additionally empirically prove that contrary to claims in previous works, although the degree distribution of these networks appear to follow a power law distribution, they are actually more in agreement with a log-normal distribution, when a numerically grounded curve-fitting is applied. Finally, we explore how syllabary and lexical networks for the English language change over time using a database of age-of-acquisition rating words. Our analysis further shows that the preferential attachment mechanism appears to be a well-grounded explanation for the degree distribution of the syllabary network.

1 Introduction

A graph is a mathematical structure that is defined by a set of vertices (or nodes) that are potentially connected by edges (or links). In the last two decades, the formal study of graphs as well as their applications have received tremendous attention from the scientific community resulting in an exponential growth of academic publications.

The rapid rise of this field of research stems from the synergy of two main factors. First, this discipline benefited from the definition and characterization of special networks such as small-world networks (Watts and Strogatz, 1998) and scale-free networks (Barabási and Albert, 1999; Albert et al., 1999). Second, the amount of digital data has doubled in size every year (Lv et al., 2017), resulting in significant developments in associating various aspects of languages with graph-based approaches (Todorovska et al., 2023; Quispe et al., 2021; Liang et al., 2019).

In this work, we investigate the use of social networks in the context of syllabification. Syllabification has been used to study the division of a word into its constituent syllables and units of pronunciation. The syllable constitutes the key building block in phonetics (Laver, 1994) and in phonological theory (Fudge, 1969; Hooper, 1972; Selkirk, 1982). It aids word modelling in automatic speech and concatenative synthesis (Marchand and Damper, 2007). For instance, Müller, Möbius, and Prescher (2000, p.225) (Müller et al., 2000) write "syllable structure represents valuable information for pronunciation systems." In morphology, syllabification is also critical to understanding word formation as well as subsequent morphological changes (Ding et al., 2019). Furthermore, syllabification has greatly contributed to the comprehension of language acquisition (Langus et al., 2017), as it helps identify the pronunciation and rhythm of words. In our work, we study syllabification using network analysis, a powerful framework for revealing structural patterns that traditional linguistic methods may overlook.

Our work explores syllable networks and lexical networks in four European languages: English, French, German and Spanish. We quantitatively examine the organization of syllables and lexicons in the languages using network analysis. We have concerns that the visual estimation of log plots for degree distribution patterns is inadequate for identifying power law. Thus, we perform a detailed empirical analysis of the degree distribution of our networks. Furthermore, we use the concept of age-

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 197–206

of-acquisition for different words, to investigate the phenomena of preferential attachment.

2 Previous Work

Syllabification using network science has been previously explored only in three languages. First, there have been efforts to develop networks of syllables for the Portuguese language (Soares et al., 2005) in which the nodes represented syllables, and corresponding edges represented the co-occurrence of syllables in words (i.e. the pair of syllables occurred together in at least one word). The authors used two datasets for their study: (1) a Portuguese dictionary with 22,064 words; and (2) the works of Machado de Assis (Caldwell, 1970). This work claimed the presence of a mechanism of preferential attachment to explain the structure of their syllabary networks. Their finding was based on log-log plots inspection of the degree distributions as well as the calculation of the power law exponent (i.e. $gamma$), whose value was compatible with the $gamma$ value range that is commonly found in this kind of network (Albert and Barabási, 2002).

Second, syllabic and graphemic (character) networks for two Chinese dictionaries have also been generated (Peng et al., 2008) previously: (1) a Putonghua dictionary (CEDICT) consisting of 21,727 multisyllabic words and 8,834 monosyllabic words; and (2) a Cantonese dictionary (CULEX) including 35,732 multisyllabic words and 5,737 monosyllabic words. Considering the segmental (base syllable) and supra-segmental (tone) features of a syllable, the authors developed three levels of syllable networks: base-syllable, tonal syllable, and Chinese-character levels. Upon visual inspection of log-log plots and comparison of $gamma$ values, it was suggested that the syllabary networks followed a power-law distribution. Upon analyzing the power law exponent of the networks, they found an increasing order in the values of gamma, thus inferring preferential attachment to be strong at the base level, stronger at the tonal level and strongest at the character level networks.

Lastly, Croatian syllabic networks have been developed (Ban et al., 2013) using two large corpora: (1) the Croatian Wikipedia; and (2) the composition of 3218 articles from Croatian blogs. They construct three co-occurrence syllable networks and one directed, weighted first-neighbour syllable (formed by connecting only neighbour syllables) network. Although a formal power law analysis was not done, the authors visualized a log-log plot for the degree distribution of the co-occurrence syllable networks to estimate a premise for power law distribution.

Considering these works, our work explores syllable networks in four main European languages: English, French, German and Spanish. In addition to syllabary networks, we also model lexical networks to reflect the importance of words in the *mental lexicon*, a concept central to psycholinguistics (e.g., Coltheart et al., 2001; Aitchison, 2012). We analyze our networks through the lens of random, small-world, and scale-free models to develop insights into the linguistic structure and cognitive processing. We also suspect that a visual inspection of log-log plots of degree distribution is insufficient to estimate the presence of a power law accurately. Instead, statistical measures like Maximum Likelihood Estimation (MLE) and goodness-of-fit tests prove to be more accurate (Goldstein et al., 2004; Clauset et al., 2009). Thus we conduct a more rigorous numerical analysis related to the degree distribution of our networks. Finally, we use a database of age-of-acquisition rating words to look into the hypothetical mechanism that is commonly used to explain a scale-free degree distribution, that is the preferential attachment (Barabási and Albert, 1999).

3 Languages and lexicons used

Four European languages (English, French, German, and Spanish) were chosen for the current work. These languages were selected due to the availability of lexicons containing marked syllable boundaries in both the spelling (written) and pronunciation (spoken) domains[1]. All these languages are from the Indo-European family and are divided between the Germanic (English and German) and Romance (French and Spanish) subgroups (Algeo and Butcher, 2013).

All entries of these lexicons which were non-words, incomplete, or contained non-alphabetic characters are excluded from the network analysis. Additionally, proper nouns as well as all instances of homophones and homographs were also removed except in cases where these entries existed only with the same syllable boundaries. In such cases, one copy of the word was kept.

[1]We used the same methodology for the two domains. Since the results were highly similar, we only report the ones that are related to the spelling domain for the sake of simplicity and readability.

4 Network Analysis

4.1 Construction

For each of the four languages under investigation, we have developed two ways to build and study the networks, namely a 'syllabary network' and a 'lexical network'. In the first representation, the nodes represent a word and two words are connected when they share at least one common syllable. In the second representation, the nodes represent each unique syllable of the language and a link is generated when two syllable nodes have at least one word in common. The dataset and code for this study are available here [2].

Figures 1 and 2 respectively show an example of lexical and syllabary network for the following short list of 8 English words: "*a–mi–a–bil–i–ty*", "*a–vail–a–ble*", "*bin*", "*cred–i–bil–i–ty*", "*in–cred–u–lous*", "*mile*", "*sim–i–lar–i–ty*", and "*sim–u–la–tion*" (the symbol '–' denotes a syllable boundary). For instance, within the lexical network, there is a link between the words "*credibility*" and "*incredulous*" as they share the same syllable, namely "*cred*". Similarly, within the syllabary network, the syllables "*vail*" and "*ble*" are connected as the word "*available*" contains them. The nodes that do not have any connections (i.e. their degree is zero) such as "*mile*" and "*bin*" are called *hermits*. Note that the largest connected subgraph is called the *giant component* whereas an *island* is defined as a connected subgraph that is not part of any larger connected subgraph.

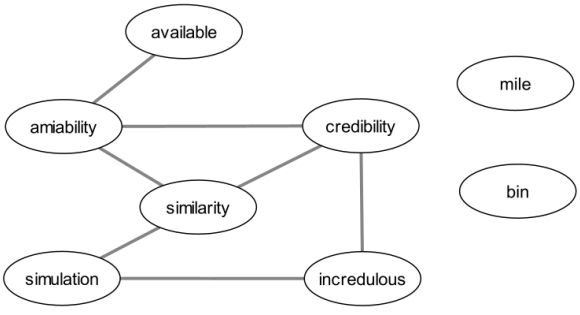


Figure 1: Example of an English lexical network

4.2 Key Properties

A network analysis was conducted by measuring a set of elementary characteristics known to formally define the main types of networks: random network (Erdős and Rényi, 1959), small-world network (Watts and Strogatz, 1998) and scale-free network (Barabási and Albert, 1999). These properties help us gain insights into the structure, cognitive efficiency and robustness of the language system. The following topological features were numerically evaluated:

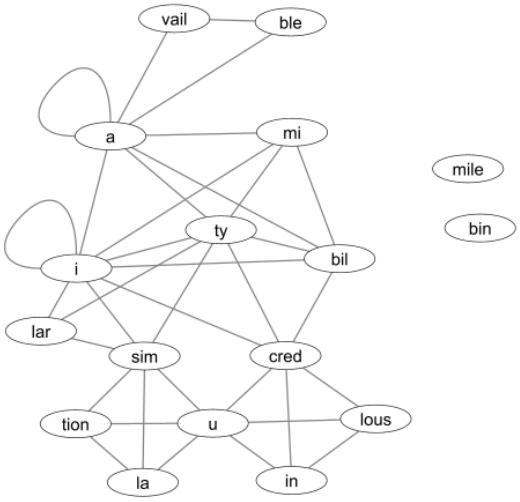


Figure 2: Example of an English syllabary network

Average Connectivity: This focuses on the degree of each node i in the network, denoted by k_i. It counts the average number of connections per node and indicates if our language networks are complex or not. If the number of nodes (N) is much larger than the average connectivity, $\langle k \rangle \ll N$, then the networks are complex by nature.

Density of a Network: Represents the proportion of possible relationships in the network that are present. It indicates the level of co-occurrences between the syllables and words. A lower density may potentially reflect phonotactic rules that restrict certain combinations.

Diameter of a Network: Defines the maximum distance between any two nodes in a network, reflecting overall connectivity, and how efficiently nodes are linked. A larger diameter for the lexical network indicates greater morphological diversity or isolated word groups. For the syllabary network, a larger diameter suggests a more fragmented structure, potentially due to rare or borrowed syllables.

Average Distance (or Average Path Length): Measures the average number of steps required to

[2] https://github.com/Rusali28/Network-analysis-syllabification-study

connect any two nodes in the network. A longer path length indicates a more fragmented language system with distinct syllable groups, thereby reflecting greater phonological diversity or complex word formation rules. A shorter length suggests that the syllables or lexicons are efficiently connected, thus speakers can easily transition between them during language processing. Small average distance (with high clustering) contributes as an indicator of small-world networks.

Average Clustering Coefficient: Measures the tendency of nodes to form tightly connected groups in the network. It determines how well a node's nearest neighbors are also connected to each other. High clustering coefficients for the lexical network suggest that words sharing syllables tend to form dense groups, reflecting phonotactic consistency and morphological relationships. High coefficients for the syllabary networks also indicate that frequently co-occurring syllables form dense clusters, revealing common phonological patterns. A high clustering coefficient is also a characteristic of small-world networks.

Distance of a Random Network (Erdös–Rényi Model): Same as the average distance. It is used to compare values with the lexical and syllabary networks, to identify their small-world nature.

Clustering Coefficient of a Random Network (Erdös–Rényi model): Similar to the average clustering coefficient. It is calculated to compare values with the lexical and syllabary networks, to identify their small-world nature.

4.3 Degree Distribution

The distribution of degree, denoted as *P(k)*, is used to measure the frequency of nodes in a network given a *k* connectivity. *P(k)* is a key metric for characterizing the structure of a network. For instance, scale-free networks have power-law distribution (Barabási and Albert, 1999) whereas random networks have Poisson-type distribution (Erdős and Rényi, 1959). To formally and adequately describe the degree distribution of the networks of our study, we used three main candidate models for curve fitting: power law distribution, lognormal distribution, and exponential distribution.

Power Law Distribution: Suggests that a small number of nodes have high connectivity, while the majority of other nodes have fewer connections. This structure is characteristic of scale-free networks, where frequently occurring syllables continue to gain more connections over time (growth driven by preferential attachment). If our networks follow a power law, it will indicate a pattern of hierarchical organization, where syllable usage is dominated by a few central syllables that appear in many words.

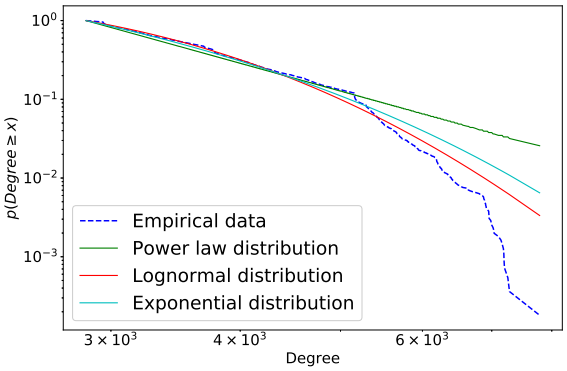
Lognormal Distribution: Suggests that, while some nodes may be highly connected, the distribution of connections is more balanced, with many nodes having moderate connectivity. Unlike a strict power law, it allows for a more gradual transition between high- and low-degree nodes. With respect to syllabification, this distribution implies that syllable connectivity is not only influenced by preferential attachment but also by phonotactic constraints and linguistic rules. This suggests that although some syllables are more common, their distribution is shaped by additional factors beyond just frequency-based reinforcement.

Exponential Distribution: Suggests that connections between nodes are relatively uniform, indicating that there are no highly dominant syllables or lexicons. This implies that syllable usage is almost random, without strong structural constraint or preferential growth. An exponential distribution in our networks will suggest that all syllables have almost equal probability of occurring in different words, which contradicts known linguistic patterns of syllable frequency and phonotactic constraints. Some previous studies (Vitevitch, 2009; Masucci and Rodgers, 2006) on language networks have suggested the presence of exponential degree distributions in certain conditions like highly restricted phonological systems. This model will help us verify whether syllabification networks follow a simple decay process rather than a structured process of phonology and lexical organization.

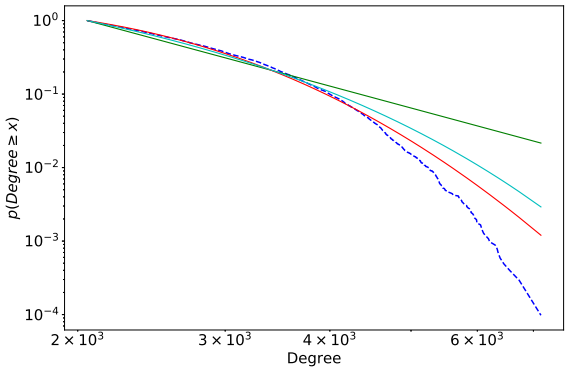
5 Results

Tables 1 show the results of the key properties for the lexical and syllabary networks, respectively. In both tables, the details of the largest connected component (i.e. giant component) are provided. The islands in our study represent small separate networks within the network consisting only of 2 to 3 nodes.

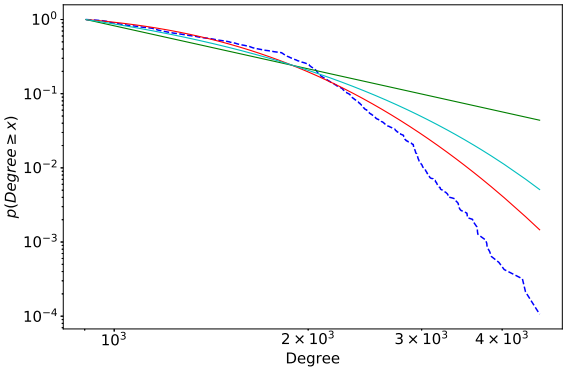
Figure 3: Word as node (lexical network).



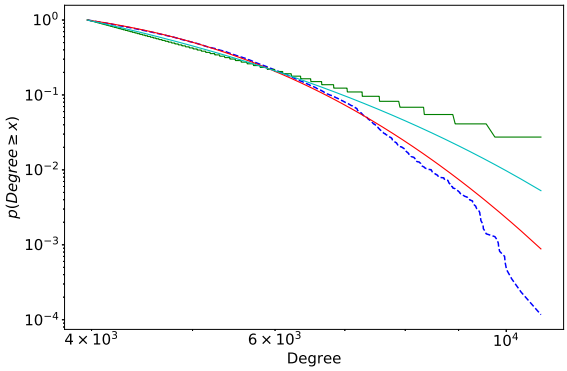
(a) English lexicon.



(b) French lexicon.

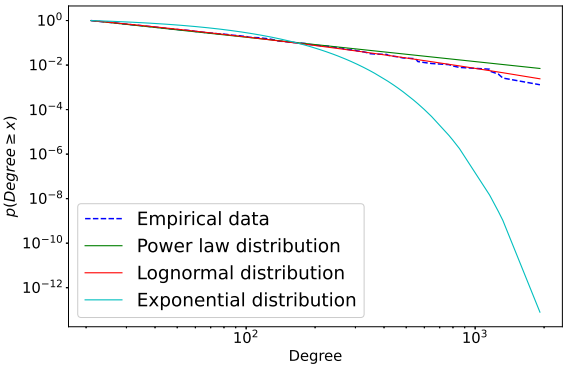


(c) German lexicon.

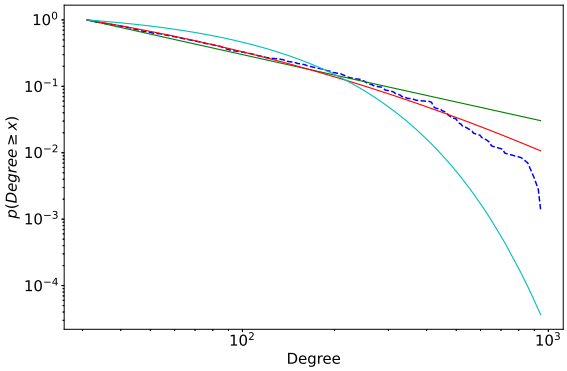


(d) Spanish lexicon.

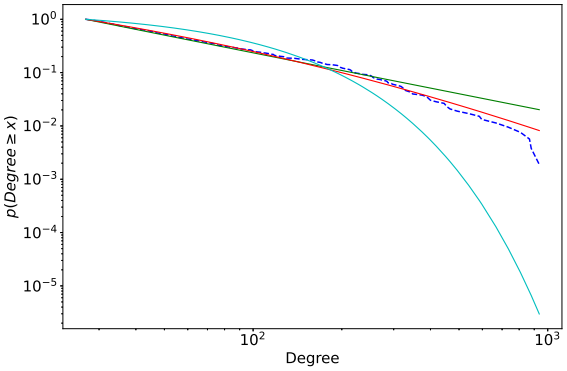
Figure 4: Syllable as node (syllabary network).



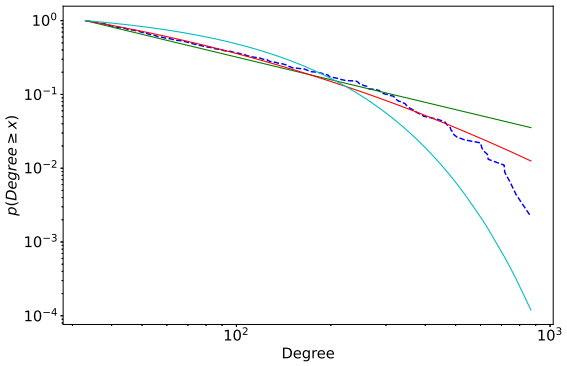
(a) English lexicon.



(b) French lexicon.



(c) German lexicon.



(d) Spanish lexicon.

Table 1: Values of the key properties for the lexical and syllabary networks.

Features	Lexical Networks				Syllabary Networks			
	English	French	German	Spanish	English	French	German	Spanish
Number of nodes	33,464	31,155	20,344	31,238	7,605	4,664	4,043	2,264
Number of islands	8	5	3	0	28	20	8	3
Number of hermits	1,245	297	84	95	1,225	282	79	93
Giant component								
Number of nodes	32,202	30,847	20,253	31,142	6,321	4,338	3,947	2,165
Number of links	20,352,425	24,235,869	9,676,660	46,186,583	44,698	53,430	35,868	34,945
Avg connectivity	1,216.4	1,555.8	951.3	2,957.1	14.1	24.6	18.2	32.3
Density	0.039	0.051	0.047	0.095	0.001	0.003	0.002	0.007
Avg distance	2.3	2.0	2.1	1.9	2.9	2.8	2.8	2.5
Diameter	9	7	7	5	10	8	7	6
Avg clustering coeff.	0.73	0.6	0.67	0.59	0.41	0.5	0.54	0.69

Table 2: Average distance and clustering coefficient for the lexical networks and their random counterparts.

Language	Lexical networks		Erdös–Rényi random networks	
	Average distance	Clustering coefficient	Average distance	Clustering coefficient
English	2.3	0.73	1.5	0.039
French	2.0	0.60	1.4	0.051
German	2.1	0.67	1.4	0.047
Spanish	1.9	0.59	1.3	0.095

The most striking difference between the four languages is the number of hermits elicited in the English language for the lexical network. It is about one order of magnitude larger than its counterparts. This finding reflects the fact that the English language has an unusually large number of loan words compared to other languages (e.g. "*zigzag*", "*seigneur*" are hermits because they are words whose syllables are unique). England being the subject to several invasions throughout its history, the English language ended up absorbing immense amounts of foreign vocabulary from Greek, Latin, Norman-French, Old French, Old Norse, and the Celtic languages as well as actively adopted words from around the parts of the world that used to be in the British Empire, and beyond (Bryson, 2001).

We observe that for all the networks, the average connectivity k satisfies the condition $\langle k \rangle \ll N$, indicating that the networks are sparse, an expected attribute for complex networks (Albert and Barabási, 2002). These results are in agreement with previous findings related to Portuguese (Soares et al., 2005), Chinese (Peng et al., 2008) and Croatian (Ban et al., 2013) language networks. Results from Tables 2 and 3 show that the networks have high clustering coefficient values when compared to corresponding Erdös–Rényi random (ER) networks whereas their average network distances are quite similar to the distances calculated for the ER random networks. Thus, our networks exhibit characteristics of small-world networks (i.e. high clustering coefficients and small average distances).

Of critical importance, we also numerically tested whether or not the empirical degree distribution of the networks was heavy-tailed by fitting three candidate distributions to the data : a power law distribution, a log-normal distribution, and an exponential distribution (Alstott et al., 2014). In the case of lexical networks, it is apparent from the four plots shown in Figure 3 that the log-normal distribution offers a better fit than the exponential model, which is not heavy-tailed. The values of the mean (μ) and standard deviation (σ) supporting the lognormal distribution are given in Table 4, for all the networks.

Concerning the syllable networks, Figure 4 shows that the degree distribution of those networks can be reasonably well modeled with a power law distribution. The associated γ values are displayed in Table 4 and are consistent with the values of γ generally observed in complex networks ($1 < \gamma < 3$). This finding is also comparable with values of complex networks that follow preferential attachment. In agreement with the study conducted by Broido and Clauset (Broido and Clauset, 2019), we also find that the lognormal distribution is a better fit for our networks than the power law distribution,

Table 3: Average distance and clustering coefficient for the syllabary networks and their random counterparts.

Language	Syllabary networks		Erdös–Rényi random networks	
	Average distance	Clustering coefficient	Average distance	Clustering coefficient
English	2.9	0.41	3.3	0.002
French	2.8	0.50	2.6	0.006
German	2.8	0.54	2.9	0.005
Spanish	2.5	0.69	2.2	0.014

Table 4: Parameters of lognormal distribution and power law distribution.

Language	English		French		German		Spanish	
Nodes of the network	Words	Syllables	Words	Syllables	Words	Syllables	Words	Syllables
Gamma (γ)	4.6	2.1	4.1	2.0	2.9	2.1	4.8	2.0
Best fit	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear
μ	7.99	-8.80	7.65	0.94	7.20	-4.76	8.38	1.14
σ	0.32	3.35	0.40	2.04	0.45	2.04	0.28	2.02

even if the latter can be seen at first sight as a good fitting curve.

To further strengthen our empirical evaluation of the network's degree distribution, we perform the degree distribution fit using the two-parameter Zipf-Mandelbrot law and compare the fit against the lognormal distribution. To determine which distribution fits the best, we measure the goodness of fit using Akaike Information Criterion (AIC) and Bayesion Information Criterion (BIC). The results from Table 5 show that for both the lexical and syllabary networks, the AIC and BIC values for the lognormal distribution are lower than those for the Zipf-Mandelbrot model. Thus, these results further indicate that the lognormal distribution consistently provides a better fit for our networks. These findings therefore suggest that syllable organization is shaped by a combination of preferential attachment, phonological constraints and linguistic rules rather than an unrestricted rich-get-richer mechanism.

6 Preferential Attachment and Age-of-Acquisition

The mechanism of preferential attachment helps to understand the growth of a complex network (Barabási and Albert, 1999; Hills et al., 2009). It is associated with the metaphoric "rich-get-richer" expression. Given a network with a pre-existing fixed number of nodes and edges, when a new node is introduced to the network, the existing nodes to which the new node will connect are selected by a preferential mechanism, i.e. the probability of selection is directly proportional to the degree of the pre-existing nodes. This means that nodes with larger degrees will have higher probabilities to attract new nodes and generate more edges, as the network evolves into a larger network with time (Barabási and Albert, 1999). Owing to preferential attachment, as well-connected nodes continue to accumulate more edges and nodes with time, this mechanism also proceeds to support the power law distribution (Stumpf and Porter, 2012), hence the scale-free nature of complex networks.

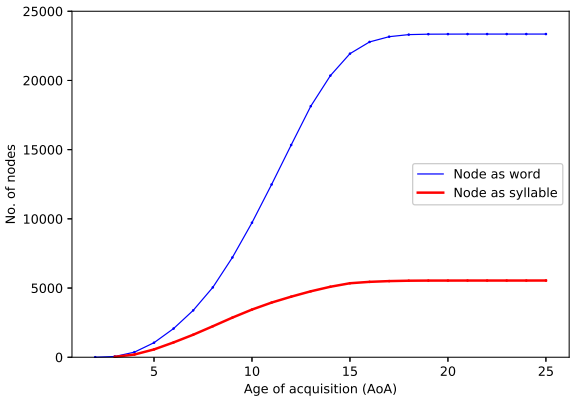
In this section, we explore the effects of an age of acquisition (AoA) database in an attempt to validate the theory of preferential attachment for the English language networks. Age of acquisition refers to the age range in which an individual acquires knowledge over certain aspects of language (e.g. vocabulary, grammar). Here, we focus on the age of acquisition of vocabulary using a dataset (Kuperman et al., 2012) of AoA ratings for 30,121 English content words including nouns, verbs, and adjectives.

We take this longitudinal vocabulary data to cumulatively construct syllabary and lexical networks over time. In other terms, a network at a time instant t is built from words known (i.e. acquired) at time t and all the proceeding words before t. Going through the AoA database, we generate 23 networks, one for each year starting from 4 to 25 years old. We perform two measures on these longitudinal networks. The first measure is the size of the networks, that is, the number of nodes in the networks. It corresponds to the measure of vocabulary size for the lexical networks and the number of syllables (an index of flexibility in word combination) for the syllabary networks. Figure 5a shows that the rate of syllabary development is slower than the lexical one. Both trajectories reach

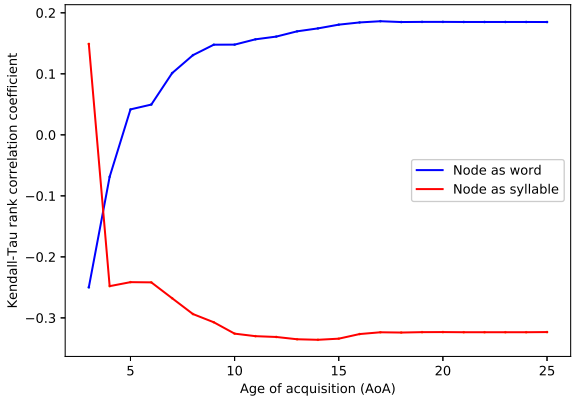
Table 5: AIC and BIC values for Lognormal and Zipf distributions in lexical and syllabary Networks.

Features	Lexical Networks				Syllabary Networks			
	English	French	German	Spanish	English	French	German	Spanish
AIC (lognormal)	91599.64	163654.51	139663.52	147018.4	7832.26	7381.2	5329.9	4731.59
AIC (zipf)	361063.23	341027.47	199321.54	374468.74	21798.45	15072.72	13185.03	7604.13
BIC (lognormal)	91612.88	163668.97	139677.82	147032.51	7841.54	7390.33	5338.44	4739.82
BIC (zipf)	361088.49	341052.51	199345.30	374493.79	21818.73	15091.88	13203.88	7621.18

Figure 5: Longitudinal analysis.



(a) Network size over time.



(b) Kendall's Tau correlation over time.

a plateau in adolescence with, at this time, a pool of about 5,000 syllables to assemble over 20,000 words.

For the second measure, we use a correlational analysis, in which the network growth allows us to assess the hypothesis of a preferential attachment. If preferential attachment[3] plays a role in shaping the English networks, then a negative relationship should be found between the AoA and degree. The first syllables that individuals acquire in earlier years should indeed be more connected over time than those learnt later in life. Figure 5b shows the Kendall's Tau correlation coefficient (Kendall, 1938) between degree and age of acquisition over time. This correlation coefficient was chosen since it is a non-parametric correlation measure that does not assume a linear relationship, thus making it more suitable for studying degree distributions, which are often skewed and not normally distributed. It provides a more robust measure of association, as opposed to the Pearson correlation coefficient (Anscombe, 1973). It can be seen that the degree and AoA of the syllabary network have a negative relationship that amplifies over time, thus supporting the theory of preferential attachment. As the age increases, newly acquired syllables tend to connect more likely to highly connected existing syllables. For the lexical network, the situation is reversed: the Kendall's Tau correlation coefficient elicits a positive correlation, reflecting different growth dynamics. As some new words are learnt over time, all their syllabic sub-parts are more likely to be linked to an existing stock of syllables. This highlights the importance of frequently used syllables, in their role as anchors in vocabulary expansion over time. The dual pattern indicates the contribution of broader linguistic constraints in shaping the overall network structure.

[3]As previously mentioned, even if a lognormal distribution is a better fit to model our networks, preferential attachment may still influence local network growth, especially in early language acquisition. Thus, power law distribution is still a reasonable fit.

7 Conclusion

We used network science to study the syllabification of the English, French, German and Spanish languages. We built syllabary networks in which nodes and links constitute syllables and words, respectively. Furthermore, to acknowledge the ubiquitous prevalence of the mental lexicon in the language sciences, we also generated and studied the properties of lexical networks in which the roles are reversed, namely nodes and links act for words and syllables, respectively. By studying the two types of networks, our study acknowledges the role of syllables in influencing word formation, as well

as the role of words in structuring how syllables interact within the lexicon. Thus, we attempt to capture both the phonological and lexical network dynamics.

All these networks exhibited small-world properties, having high clustering coefficients and small average distances when compared to their corresponding random networks. Previous studies on syllabary networks related to Portuguese, Chinese and Croatian languages, have all assumed that they are scale-free as their degree distribution looked as to follow a power law. However, this statement was mostly based on visual inspection. When an analysis of degree distribution is conducted under more rigorous circumstances using curve fitting, the results show that all our networks' degree distributions are consistently better modelled with a log-normal distribution. Our finding is in line with a few studies that have questioned and rebutted the claim of previously reported scale-free networks (e.g. Clauset et al., 2009; Broido and Clauset, 2019).

Finally, unlike earlier studies on syllabification that all dealt with static networks, we examined the English networks from a dynamic perspective using longitudinal data from the database of age-of-acquisition rating words. This approach allowed us to validate the preferential attachment mechanism for the syllabary network. Although our analysis shows that syllabification networks are better described by a lognormal distribution rather than a strict power law, this does not rule out preferential attachment as a contributing mechanism (Hills et al., 2009; Vitevitch, 2008). We find that preferential attachment can still operate locally in early network growth. This implies that language network growth is influenced by both preferential attachment and additional linguistic constraints, like phonological and morphological structures.

These findings in our work have important implications for linguistic modeling. Considering phonotactic constraints and linguistic rules have a role to play in the networks, models of lexical evolution must account for structured constraints beyond simple preferential attachment. Furthermore, cross-linguistic comparisons of degree distributions could further help us understand how these constraints vary across mutliple languages, thereby providing deeper insights into the universality of phonological organization.

8 Limitations

While our study provides new insights into syllabification networks, we address a few limitations in this section. First, the focus of our study is limited to four Indo-European languages. This generates a scope for cross-linguistic validation with typologically diverse languages, with different structures (e.g. Turkish, Finnish). Second, while our results demonstrate that a lognormal distribution better models the degree distribution than power law, we still observe reasonable behaviour with respect to power law as well. Future work could explore alternative mixture models to refine this finding. Third, while we confirm that preferential attachment influences network growth, our study does not quantify its relative impact compared to linguistic constraints like phonotactic rules, morphological constraints, cognitive biases, etc. A quantitative comparison of these factors could provide deeper insights into their contributions.

References

Jean Aitchison. 2012. *Words in the mind: An introduction to the mental lexicon*. John Wiley & Sons.

Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

Réka Albert, Hawoong Jeong, and Albert-László Barabási. 1999. Diameter of the world-wide web. *Nature*, 401(6749):130–131.

John Algeo and Carmen A Butcher. 2013. *The origins and development of the English language*. Cengage Learning.

Jeff Alstott, Ed Bullmore, and Dietmar Plenz. 2014. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1):e85777.

Francis J Anscombe. 1973. Graphs in statistical analysis. *The american statistician*, 27(1):17–21.

Kristina Ban, Ivan Ivakić, and Ana Meštrović. 2013. A preliminary study of croatian language syllable networks. In *2013 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1296–1300. IEEE.

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Anna D Broido and Aaron Clauset. 2019. Scale-free networks are rare. *Nature Communications*, 10(1):1017.

Bill Bryson. 2001. *The Mother Tongue: English and How it Got that Way*. William Morrow Paperbacks.

Helen Caldwell. 1970. *Machado de Assis: the Brazilian master and his novels*, volume 2. Univ of California Press.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. 2001. Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108(1):204.

Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2019. Towards burmese (myanmar) morphological analysis: Syllable-based tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–34.

Paul Erdős and Alfréd Rényi. 1959. On random graphs. *Publicationes Mathematicae*, 6:290–297.

Erik C Fudge. 1969. Syllables. *Journal of Linguistics*, 5(2):253–286.

Michel L Goldstein, Steven A Morris, and Gary G Yen. 2004. Problems with fitting to the power-law distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 41:255–258.

Thomas T Hills, Mounir Maouene, Josita Maouene, Adam Sheya, and Linda Smith. 2009. Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science*, 20(6):729–739.

Joan B Hooper. 1972. The syllable in phonological theory. *Language*, pages 525–540.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990.

Alan Langus, Jacques Mehler, and Marina Nespor. 2017. Rhythm in language acquisition. *Neuroscience & Biobehavioral Reviews*, 81:158–166.

John Laver. 1994. *Principles of phonetics*. Cambridge university press.

Wei Liang, Guanrong Chen, and Zihan Zhang. 2019. Adjacency spectra of chinese character co-occurrence networks in different historical periods. *Physica A: Statistical Mechanics and its Applications*, 536:122541.

Zhihan Lv, Houbing Song, Pablo Basanta-Val, Anthony Steed, and Minho Jo. 2017. Next-generation big data analytics: State of the art, challenges, and future research topics. *IEEE Transactions on Industrial Informatics*, 13(4):1891–1899.

Yannick Marchand and Robert I Damper. 2007. Can syllabification improve pronunciation by analogy of english? *Natural Language Engineering*, 13(1):1–24.

A. P. Masucci and G. J. Rodgers. 2006. Network properties of written human language. *Physical Review E*, 74(2):026102.

Karin Müller, Bernd Möbius, and Detlef Prescher. 2000. Inducing probabilistic syllable classes using multivariate clustering. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 225–232.

Gang Peng, James W Minett, and William S-Y Wang. 2008. The networks of syllables and characters in chinese. *Journal of Quantitative Linguistics*, 15(3):243–255.

Laura VC Quispe, Jorge AV Tohalino, and Diego R Amancio. 2021. Using virtual edges to improve the discriminability of co-occurrence text networks. *Physica A: Statistical Mechanics and its Applications*, 562:125344.

Elisabeth Selkirk. 1982. The syllable. *The structure of phonological representations*, 2:337–383.

M Medeiros Soares, Gilberto Corso, and LS Lucena. 2005. The network of syllables in portuguese. *Physica A: Statistical Mechanics and its Applications*, 355(2-4):678–684.

Michael PH Stumpf and Mason A Porter. 2012. Critical truths about power laws. *Science*, 335(6069):665–666.

Ana Todorovska, Hristijan Peshov, Ivan Rusevski, Irena Vodenska, Lubomir T Chitkushev, and Dimitar Trajanov. 2023. Using ml and explainable ai to understand the interdependency networks between classical economic indicators and crypto-markets. *Physica A: Statistical Mechanics and its Applications*, page 128900.

Michael S Vitevitch. 2008. What can graph theory tell us about word learning and lexical retrieval?

Michael S. Vitevitch. 2009. What can network science tell us about phonological similarity? *Speech Communication*, 51(1):5–13.

Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442.