Syllabic Networks: measuring the redundancy of associative syntactic patterns

Bradly Alicea
Orthogonal Research
bradly.alicea@outlook.com

Keywords: natural language evolution, natural language structure, networks, self-organization, information theory

ABSTRACT

The self-organization and diversity inherent in natural and artificial language can be revealed using a technique called syllabic network decomposition. The topology of such networks are determined by a series of linguistic strings which are broken apart at critical points and then linked together in a non-linear fashion. Small proof-of-concept examples are given using words from the English language. A criterion for connectedness and two statistical parameters for measuring connectedness are applied to these examples. To conclude, we will discuss some applications of this technique, ranging from improving models of speech recognition to bioinformatic analysis and recreational games.

Introduction

According to Shannon and Weiner (1949), we should expect a certain degree of redundancy to exist in any natural language. This redundancy constitutes a majority of elements in the syntax of a given groups of words, and is inversely related to the number of meaningful elements (Patterson, 1987). Such a power law relationship among elements suggests that language networks are both small-world and scale-free (Newman, 2003). In the case of one-dimensional syntactic networks, these network properties hold and may ultimately play a role in the structure and evolution of natural languages (Liu and Hu, 2008).

The structure of the syllabic network is similar to Jumbo, a program developed by Douglas Hofstadter (1995) to sort scrambled anagrams into recognizable words. According to Hofstadter (1995), Jumbo's architecture produces clusters that are analogous to complex molecules or human relationships that appear in nature residing in a chaotic environment. This results in adaptive, multi-layered structures joined by bonds of different time-dependent strengths (Hofstadter 1995). In this case, the generated strings have both "hinge points" and "breaking points". Some subdivisions are more natural while others could take on a number of different forms given a minimum amount of structural adjustment.

The formation of syllabic networks rely on words being chunked into several portions. While these chinks are not technically syllables, the arbitrary nature of the chunking process is similar to how information is processed during perception (Miller, 1956). Dynamic chunking reduces the complexity of linear sequences across natural languages (Lu, Xu, and Liu, 2016). According to Harris (1995), the morphological structure of language can be analyzed by investigating how language consists of separate

encapsulated grammatical systems for different lists of words. A syllabic network can uncover the general relatedness between words based strictly on structure and syntax. While these structural similarities between words often have nothing to do with meaning (Harris, 1991), the kinds of creative wordplay that emerge from wordlists used to create the network topologies often have the quality of semantic free-association.

Methodology

Syllabic networks require a list of words or character strings with the following characteristics: 1) being composed of a finite alphabet with discrete symbolic states, 2) a set of initial conditions, and 3) a set of directly and/or indirectly shared characters. Theoretically, these networks can be created from any set of strings in a symbolic or natural language. In this paper, only cases from a written natural language (English) were tested. Two statistical parameters were developed to measure granularity and pattern density within and among these networks. The average granularity, or *G*, can be defined mathematically as

$$G = \frac{C}{N}$$

where *C* is the number of characters per node, and *N* is the total number of nodes. Since most network topologies will be scale-free, the number of characters per node will show significant variability. Pattern density, or *D*, is defined as

$$D = \frac{N}{S}$$

where *N* is the total number of nodes, and *S* is the number of recognizable strings contained in the network. Any number of strings can be used to build a network. Yet the selection of very few or very large numbers of strings at random may not ensure the construction of a continuous topology.

Results

We can construct syllabic networks by selecting a series of word strings and then breaking them into syllabic units shared amongst two or more words. These units can then be linked together to form whole words While some words will share syllables, not all words will have syllables in common. As will be shown, such patterns of connectivity can result in a number of topologies. In the case of the English language, strings and patterns are words defined by the discrete symbols of the Roman alphabet, but any language that utilized an alphabet with a finite number of characters will suffice.

It can be shown that a syllabic root can exist in a syllabic network topology which is common to several words. For example, the substring "vio" in the left frame of Figure 1 is part of the words violin, violence, and violate. In turn, the substring "late" is also part of the word prelate. The right-hand frame in Figure 1 demonstrates that several derivative network topologies can result from the same set of strings.

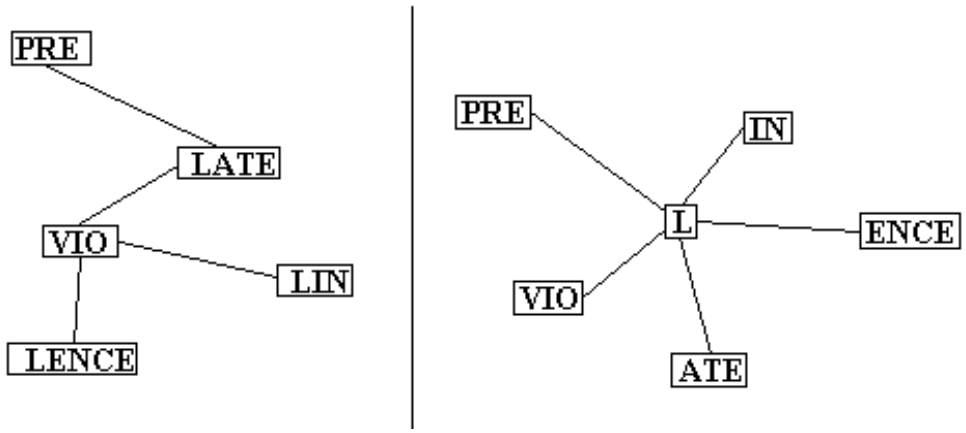

Figure 1: Two examples of a syntactical web linking multiple words to common syllabic roots.

The statistical parameters discussed previously can be directly applied to each syntactical web produced. In addition, the number of strings can be compiled by counting the number of words used in constructing the topology. This can be used for comparative purposes, or simply to track the number of strings contained in a certain category. Figure 2 shows two different syntactical nets with different sets of words used for each. We can also compare the average granularity and density parameters.

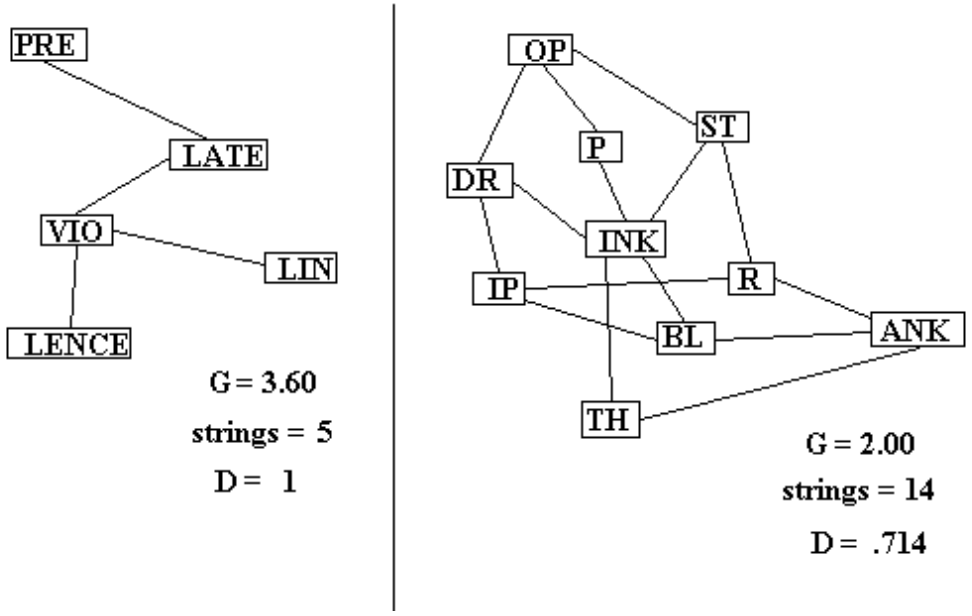

Figure 2: Two syntactical webs including both statistical parameters and the number of constituent strings.

In particular, Figure 2 shows that when fewer strings are used, the average granularity is higher and the density parameter is lower. This should not be taken as a rule, but rather that in general, higher granularities in networks with fewer strings result in higher densities. However, the occasional use of single character nodes to link together a great

many words can result in a higher density. In fact, when networks consisting of only a single character per node were tested, they yielded a density of one.

Types of Connectedness

Different word patterns can coexist in either a nested, overlapping, or discontinuous relationship with each other. Table 1 demonstrates the relative frequencies of each type of relationship for one of the topologies shown in Figure 2. These frequencies for each instance are derived by dividing the number of occurrences of a specific instance by the total number of instances.

Table 1: Matrix denoting the frequency of relationship types for the network in the left frame of Figure 2.

	Prelate	Late	Violate	Violence	Violin
Prelate	---	N	O	D	D
Late	N	---	O	D	D
Violate	O	O	---	O	O
Violence	D	D	O	---	O
Violin	D	D	O	O	---

In Table 1, there are twenty unique relationships between pairs of words. Two of these instances are nested (N), ten of them overlaps (O), and eight are discontinuous (D). Alternately, the frequency of instances is .1 for nested patterns, .4 for discontinuous patterns, and .5 for overlapping patterns. Generally, when the frequency of overlapping and/or nested patterns is closer to one, the more dense with patterns a syntactical web becomes. These type frequencies should be used in tandem with the other statistical parameters previously discussed.

Syllabic networks can yield all possible relationships between word chunks and more generally demonstrate interconnections between words. This is done by treating the network topology as a perfect lattice and making new linkages where they do not already exist. These potential linkages can be responsible for creating an infinite number of strings which have no semantical context in the language being analyzed. These potential strings provide a rough clue as to the syntactical diversity of a given language. Furthermore, recurrent patterns and relationships between nodes and their contents can be more easily discovered.

In the topology shown in Figure 3, all possible connections are made. The solid lines represent active links, while the dashed lines represent potential linkages which do not form real patterns. Thus, there can be either active or null patterns in a given topology. A given class of these null patterns thereby constitute a null structure. For example, the string "black" is an active pattern and the link between "b" and "l" is an active link. By the

same token, the string "blaret" exemplifies a null pattern, and the link between "r" and "et" is a potential linkage. The classification of linkages and patterns are also dependent on context, so that what is potential in one instance is actual in another.

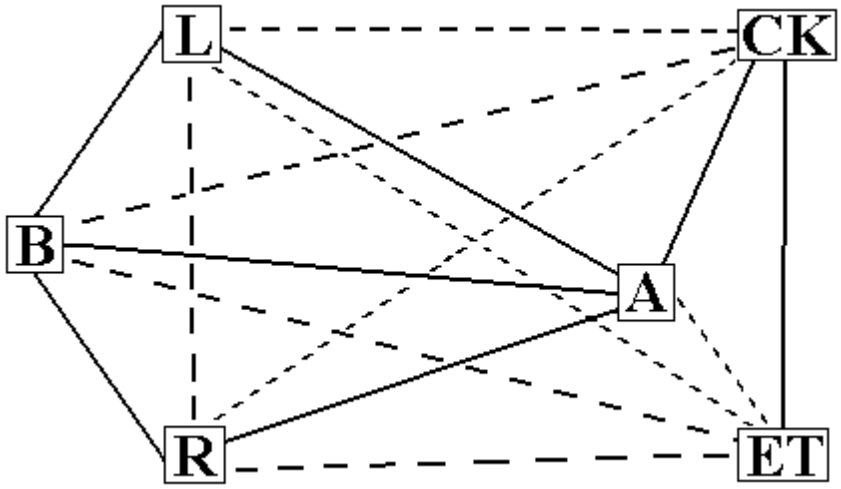

Figure 3: Syntactical network with all possible connections made between nodes.

Discussion

The topological diversity observed in syllabic networks occur in part because of the recurrent use of phonemes in speech (Ladefoged, 1975). The topology produced among these networks also uncovers syllabic patterns that repeat within a single language or linguistic family. In other words, some letter combinations may be more or less highly connected when dealing with random words from a single language. Syllabic networks also share similarities with syntactic dependency networks (Cech, Macutek, and Zabokrtsky, 2011; Mehler et.al, 2016), although the information extracted is unique for each type of network.

Syllabic networks may also have practical applications. Voice synthesizers and voice recognition software use recursive transition networks (RTNs) to synthesize sentences and entire bodies of coherent text (Bulhak, 1996). RTNs rely on stringing together syntactic chunks to form sentences, and increasing our understanding of syllabic connectivity might contribute to their continued development. Given their independence of semantic context, syllabic networks can also be used to analyze patterns and discover other shared characteristics amongst multiple DNA and protein sequences. Finally, syllabic networks can be used to analyze the syllabic and single-character connections between whole words. This is true for a number of types of relatedness, including as the contents of a crossword puzzle (sharing single letters) or as the content of a freely-associated list of words.

References

Barabasi, A.L. (2001). Linked: the new science of networks. Perseus Press, Cambridge, MA.

Bulhak, A.C. (1996). On the simulation of Postmodernism and mental debility using recursive transition networks. Department of Computer Science, Monash University, Melbourne, Australia.

Cech, R., Macutek, J., and Zabokrtsky, Z. (2011). The role of syntax in complex networks: Local and global importance of verbs in a syntactic dependency network. Physica A 390 (2011) 3614–3623.

Harris, Z.S. (1995). Theory of Language and Information: a mathematical approach. Clarendon Press, New York.

Hofstadter, D. (1995). Fluid Concepts and Creative Analogies: computer models of the fundamental mechanisms of thought. Basic Books, New York.

Ladefoged, P. (1975). A course in phonetics. Harcourt Brace Jovanovich, New York.

Liu, H. and Hu, F. (2008). What role does syntax play in a language network? Europhysics Letters, 83(1), 18002.

Mehler, A., Lucking, A., Banisch, S., Blanchard, P., Job, B. (2016). Towards a Theoretical Framework for Analyzing Complex Linguistic Networks. Springer, Berlin.

Miller, G.A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review, 63(2), 81-97.

Newman, M. (2003). The Structure and Function of Complex Networks. SIAM Review, 45, 167-256.

Patterson, W. (1987). Mathematical Cryptology: for computer scientists and mathematicians. Rowman and Littlefield, Totowa, NJ.

Shalloway, A. and Trott, J. (2001). Design Patterns Explained: a new perspective on object-oriented design. Addison-Wesley, Boston.

Shannon, C. and Weiner, N. (1949). Mathematical Theory of Communication. University of Illinois Press, Champaign, IL.