Time Driven Analysis of Synonym Networks in the English Language

Minymoh Anelone, Megan McAndrews, and Howard Small

Introduction:

In this paper, we explore several questions about the development of the English language using Princeton's WordNet (3.1) graph, which connects a large portion of the words in the English lexicon by synonymy and hyponymy. In conjunction with historical data from the Oxford English Dictionary, we examine how the language has changed in the context of these three primary questions:

1. Are older words more central?
2. How do branching factor and speed vary based upon the age of a word?
3. How do parts of speech differ by branching factor and speed?

These are important questions because they add to our body of knowledge about how English has developed, as well as the role of different parts of speech and of words' influence over time. They were difficult to answer before because they require both synonym and historical data, and the integration of both into one network. Perhaps for this reason, few have looked into WordNet from a historical perspective. Additionally, the questions require detailed definition of different measures and complex null model development.

Related Work:

The most relevant associated work is the paper "The Large Scale Structure of Semantic Networks"[1] by Mark Steyvers and Josh Tenenbaum. Steyvers and Tenenbaum analyze 3 different types of semantic networks (including WordNet), and show that those networks have small-world structures. The authors then go on to propose a model for the growth of semantic networks which generates expected small world statistics and also predicts correlation between a word's age-of-acquisition and its connectivity.

While Steyvers' and Tenenbaum's work gave us helpful context and a point of comparison, their aims differed from ours in a couple of key ways. They focused more on the dynamic evolution of the graph, focusing on measures such as age-of-acquisition by humans, while we wanted to look at the entire graph in retrospect in the context of aspects like parts of speech and year of entry into the language. Additionally, since it was so difficult to scale the growing network model, their final model included 5,018 words. Consequently, it is difficult to extrapolate the implications for an entire language. Our network, on the other hand, had around 60,000 words, even when including only those words for which we had time data.

[1] Steyvers, Mark, and Josh Tenenbaum. "The Large-Scale Structure of Semantic Networks." http://psiexp.ss.uci.edu/research/papers/smallworlds.pdf. 10 Dec. 2016.

Graph Models:

Model Data:

For our project we made use of two datasets: 1) The WordNet graph data from wordnet.princeton.edu and 2) data we scraped from the Oxford English Dictionary (OED) website with the year of the first recorded use of every word.

At a high level, WordNet is structured as a collection of synonym sets (synsets) which each contain a number of words grouped together by a common meaning. Connecting these synsets are two types of connections: connections between *entire* synsets and connections between individual words in *separate* synsets. The WordNet data also contains information about the "types" of connections in the graph (e.g. synonym, antonym, hypernym, meronym, etc.), however, we decided not to include these distinctions in our analysis because we were interested in the general connectedness/influence of words and not the specific types.

In order to do our analyses, we constructed a number of different WordNet graph models which connect the words based upon different rules. Each of the models is described below:

Model #1 (The "Standard" WordNet Graph):

To construct our standard model, we started by creating a "supernode" to represent each synset. Then for each word in a synset we created a regular node, which we connected to the supernode and all of the other nodes in the synset. Thus, every synset is present in the network as a completely connected cluster in the graph with a representative supernode. Once all of these initial connections were made, we added in all of the synset-to-synset connections as edges between the related supernodes, and the connections between words in separate synsets as edges between the nodes for the specific words. Every connection in this graph was undirected. This model is similar to the model used for the WordNet graph in the Steyvers and Tenenbaum paper.

Model #2 (The "Time Directed" WordNet Graph):

Since we were interested in the development of the structure of the English language over time, we also constructed time directed version of the WordNet Graph. Using the first-year-of-occurrence data collected from the OED, we constructed a directed version of the WordNet Graph using the method described for Model #1 with a minor alteration. When creating the synset clusters, word nodes are connected using a direct edges from the older word to the newer word.

It is important to note that the time directed graph is smaller than the original WordNet graph because of the limited overlap between words and phrases in WordNet and words and phrases with viable time data from the OED. The number of words in this graph was ~60,000.

Model #3 (The Augmented "Time Directed" WordNet Graph):

In order to more fully capture relations between words, we constructed a third model of the WordNet Graph that excluded supernodes. In the previous two models, it

is possible for supernodes to be connected without words in each of the separate synsets being connected. Since supernode connections imply a relationship between words in each of the synsets they represent, we constructed a graph that captured those relationships as well. This third model was a weighted directed graph that contained only word nodes, where words in the same synset were connected with weight 1, words that were directly connected in the previous model were connected with weight 1, and words in synsets whose supernodes were connected were connected with weight 1/2. These weights are reflective of the fact that words in the same synset and words with direct relations are more strongly related than words whose only relation is through their supernodes.

Null Models:

To further analyze the properties of our different models, we constructed randomized null models for each graph. The basic process for doing this was to randomly shuffle all of the synset-to-synset connections and also randomly shuffle the between-synset connections in the graph. This allowed us to maintain the "local" structure of the graph, with clusters for synonyms, but to mix up the "global" structure of the graph effectively. Additionally, we maintained the year data associated with every word because we felt shuffling that data would erase any of its meaning.

Graph Algorithms:

Two of the key questions in our project involved determining the branching factor and branching speed of words in our network. Here, we formally define what each of these metrics mean, and how they are computed. Both of these algorithms were designed for and run on our Model #3 graph.

Branching Factor: Branching factor is intended to represent the "reach of influence" for a given word in the graph. To calculate this, we started from the node associated with the input word and traversed along every out-going edge from this node, adding the weights of these edges to the total. Then, from each of the neighboring nodes along these outgoing edges, we repeat this process up to a depth d (a parameter for the algorithm), but at each step, the added weight of the edges is weighted by the product of the previous edge weights on the path. This additional weighting is included to make sure that traversing along a "less important" connection in the graph influences the branching factor less.

Branching Speed: Branching speed is intended to represent how quickly a word branches into new words (in time). To calculate this, we first find the influence set of the node associated with the input word in the graph. In other words, this is all nodes which can be reached from the input node within d (a parameter for the algorithm) steps. Secondly, we compute the average age of the the words younger than the input word. Note that this is the *expected* average year of the words in the influence set of the input word, if the connections were distributed randomly. Branching speed is then computed as follows:

$$speed = 1 - \frac{average\ year\ of\ words\ influence\ set - year\ of\ word}{expected\ average\ year\ of\ words\ in\ influence\ set - year\ of\ word}$$

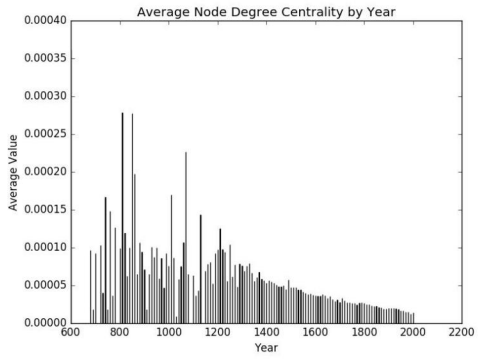
This means that words with a branching speed of zero branches to new words as expected, a word with a positive branching speed branches into new words faster than expected, and a word with a negative branching speed branches into new words slower than expected.

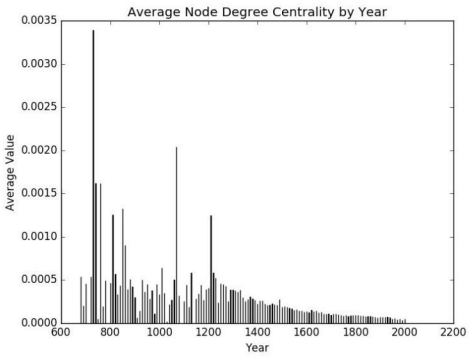
Results:

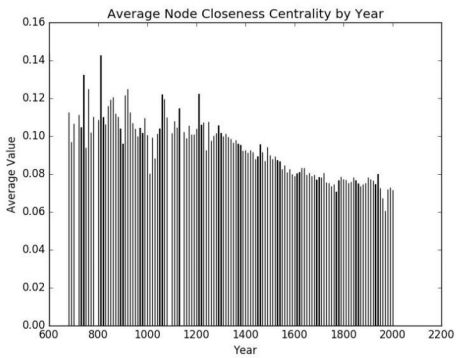
Are older words more central?

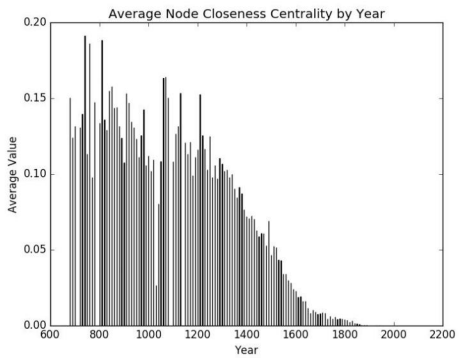
We found that older words *are* more central in the WordNet graph. While this trend held for both the time directed graph and the augmented time directed graph, it was much more pronounced in the augmented time directed graph. This result may be due to supernodes in the time directed graph adding noise during the centrality calculations. We see that for all centrality calculations, there is a trend towards older words being more central. These results are expected given the way in which the WordNet graph evolves. In terms of degree centrality, we expect older words to have higher degree since they have both a larger possible impact set as well as more time to influence words in the language. For both closeness centrality and betweenness centrality, we expect older words to have a higher centrality since the WordNet graph grows naturally outwards, so earlier words will be closer to the center of the graph.

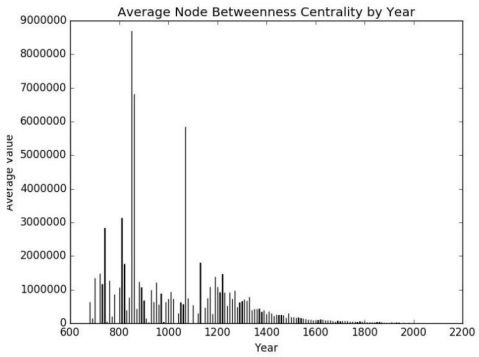
Time Directed Graph *Augmented Time Directed Graph*

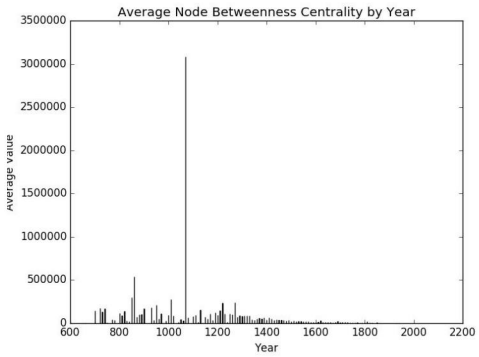








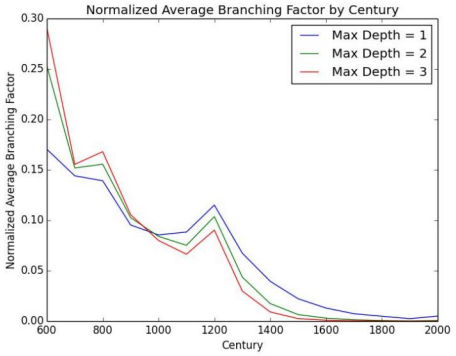


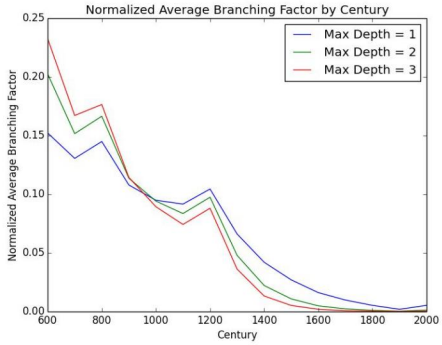


How do branching factor and speed vary based upon the age of a word?

Note: All of the graphs in this section show results from our Augmented Time Directed Graph (on the left) and the associated null-model (on the right).

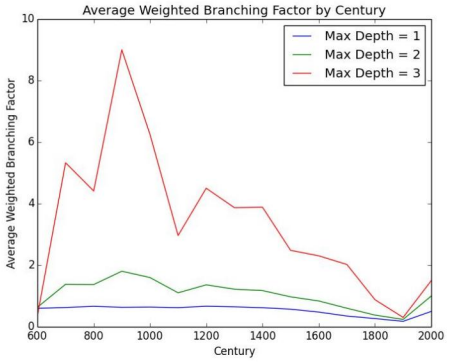
We can see from the graphs below that there is a very distinctive downward trend in branching factor as we move towards more recently introduced words. This is exactly what we expected, given our hypothesis that older words have a larger influence on the English language. Interestingly, we see that changing the depth of the branching factor calculations did not have much impact on the relative averages of the branching factor, except for words between 600-700. This implies that older words tend to have a further reaching impact in the graph than newer words. When we look at the graph for the null model, we see that downward trend still holds, which means that what primarily affects branching factor is the local, synonym-related connections. This is a surprising result because we would have expected that having connections to other parts of the graph to play a much larger role in the branching factor.

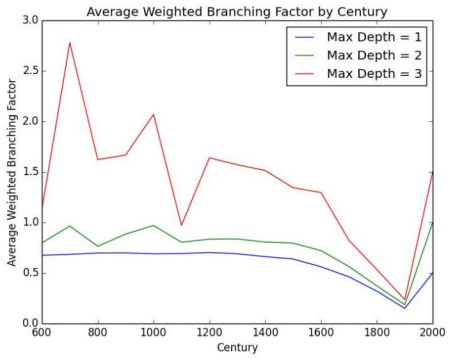




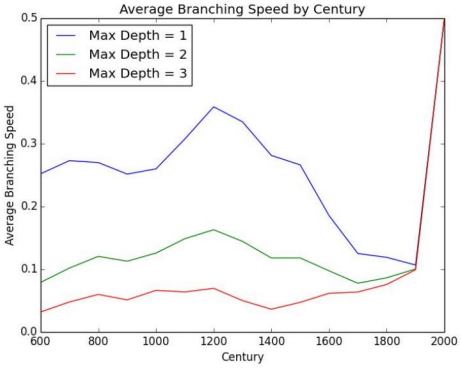
One problem with the above graphs is the inherent bias built into our network for older words. Because all connections in the graph move forward in time, earlier nodes are biased to have a higher out degree, as we can see in the degree centrality graphs above. To account for this, we modified our metric by weighting the branching factor for a given node by $(\text{out degree})^{\text{depth}}$. This metric produced the following graphs, where we can see that there is still a downward trend in the weighted branching factor over time.

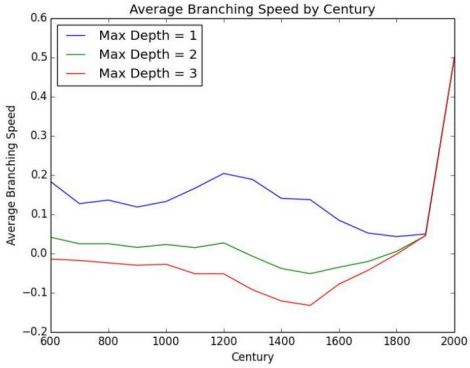
This shows that older words have a further reaching influence in the graph than newer words.





When we look at the graphs for branching speed (below), we see a more complicated trend. For depths 1 and 2, the branching speed increases until the 1200s and then decreases until the 1900s, whereas for depth 3, the branching speed increases more or less consistently, albeit marginally, throughout time. In these particular graphs, the data for the 2000s is omitted because there are so few words in the graph at that age that the branching speed calculations break down. These trends are interesting because it implies that there was a large spike in new words that quickly influenced new words and died off in the 1200s. We hypothesize that this may be related to the fact that the oldest manuscripts we have of Middle English date back to 1150[2], so there are likely a number of words that were first introduced in this time, and these words are also likely connected with one another since they were introduced in similar times. Also, the Magna Carta was written in 1215, which similarly explain the spike in words. Overall though, we see that there is not much of a universal trend for branching speed. Looking at the null model graph, we see that there is even less of a trend over time, implying that there is more to learn about how the branching speed of words relates to their year of acquisition.





[2] http://www.bl.uk/learning/langlit/evolvingenglish/accessvers/1100s/index.html

How do parts of speech differ by branching factor and speed?

We found significant differences in branching factor and speed among different parts of speech. First, in comparison with our null models, we found higher branching factors and speeds for every part of speech at every depth, with minimal exception (see Tables 1 and 2). There was also a consistent trend of higher depths reporting higher branching factors and speeds.

Table 1. Branching Factor by Part of Speech for d = 1, 2, 3

Part of speech (sample size)	No Supernodes	Null Model
Adj (11734)	2.87, 15.3, 95.5 Weighted: 0.871, 1.51, 4.74	2.79, 11.8, 46.3 Weighted: 0.729, 0.901, 1.48
N (28173)	2.76, 16.1, 97.3 Weighted: 0.691, 1.30, 3.76	2.40, 9.55, 37.1 Weighted: 0.704, 0.912, 1.59
Adv (509)	9.43, 114, 1138 Weighted: 0.806, 1.19, 2.80	9.10, 67.4, 368 Weighted: 0.715, 0.688, 0.913
V (10149)	7.85, 87.1, 835 Weighted: 0.725, 1.17, 3.34	6.97, 46.9, 250 Weighted: 0.706, 0.767, 1.37

Table 2. Branching Speed by Part of Speech for d = 1, 2, 3

Part of speech (sample size)	No Supernodes	Null Model
Adj (11734)	0.191, 0.119, 0.0835	0.116, -0.0122, -0.0644
N (28173)	0.175, 0.0857, 0.0471	0.0899, -0.0119, -0.0518
Adv (509)	0.228, 0.0947, 0.0336	0.105, -0.0341, -0.0937
V (10149)	0.308, 0.147, 0.0743	0.113, -0.0236, -0.0804

Among different parts of speech, there was variation in both branching factor and speed. For unweighted branching factor, adverbs and verbs had higher branching factor than nouns or adjectives (which had very similar branching factors), and the difference was especially pronounced at higher depths. When we weighted the branching factor to account for the increased out-degree of older nodes, we found that adjectives actually had the highest branching factor, followed by nouns and verbs, with adverbs having the lowest branching factor. This is shown in Figures 1 and 2.

Figure 1

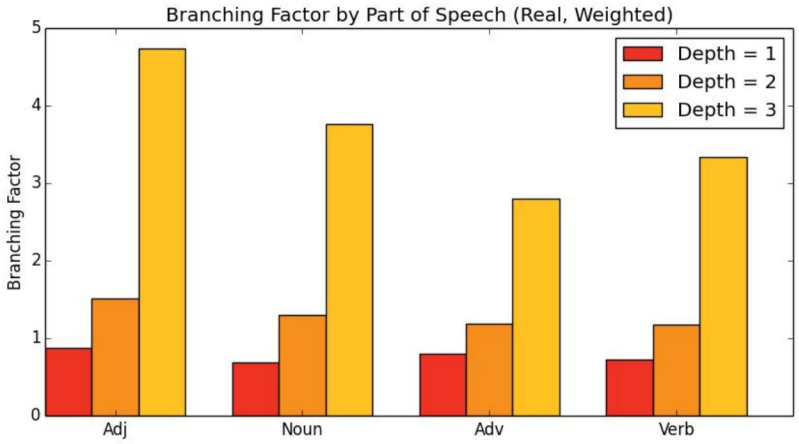
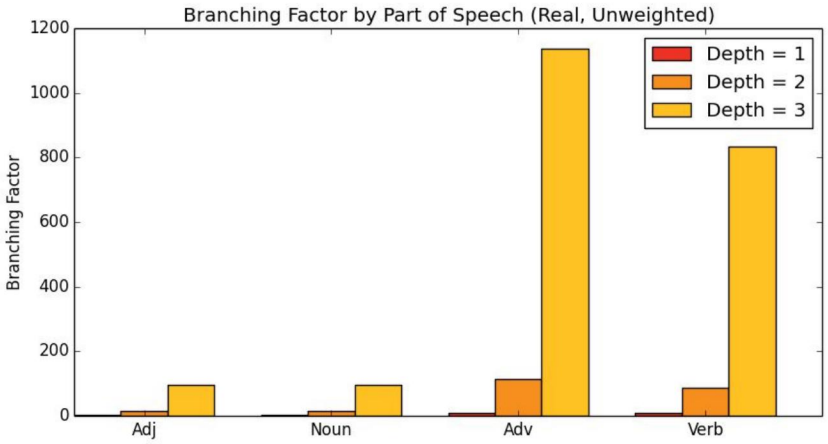


Figure 2



These results were interesting to compare to our null model results (see Figures 3 and 4); though our unweighted results were similar between the null and real models, our weighted results showed less variation among the parts of speech. This points to notable results in terms of different parts of speech showing different branching factors in WordNet. The results that adjectives have the highest branching factor makes sense given that adjectives are words meant to describe other words and would likely have more nuance and more related words. On the other hand, nouns and verbs themselves describe a single object or action, making them less likely to have an elevated number of synonyms.

Figure 3

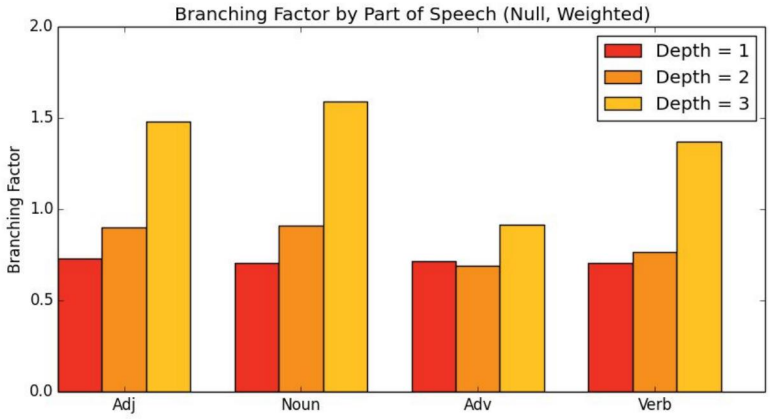
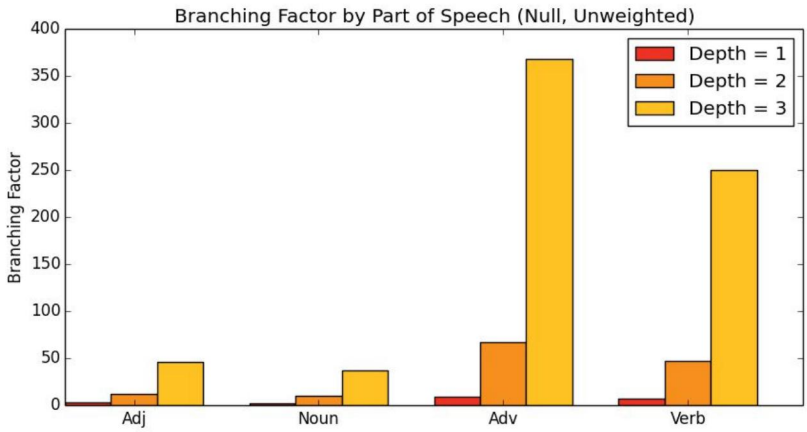


Figure 4



Looking at branching speed also provided some interesting results. At a depth of 3, adjectives and verbs had the highest branching speed, followed by nouns and finally adverbs (see Table 2). Given that higher branching speeds correlate with words having a shorter, more concentrated influence, and then fading, this would suggest that nouns have a longer-lasting influence than adjectives or verbs. This is an interesting result, particularly given that it differs from our null model result for branching speed. It may be due to the fact that once a noun describes an object, that object does necessarily require or encourage other descriptors. On the other hand, adjectives and verbs might encourage more variety and consequently change at a more rapid pace. It is worth noting that adjectives, which have the highest branching speed, are used to modify nouns; thus when nouns require nuance, it could actually increase the branching speed of adjectives rather than nouns.

Conclusions:

In this paper, we assembled a network encompassing WordNet synonymy data and OED historical data. Upon analyzing these networks, we demonstrated that older words are more central, that older words have a significantly higher branching factor but don't differ much in speed, and that adjectives have the highest branching factor and speed among parts of speech. Interdisciplinary future directions worth pursuing could include a study of whether children's language acquisition mirrors the development of the English language over time (that is, if children learn older words first), or looking at how literature across time periods captures the branching trends that we found based on time period and part of speech.