An Investigation on the Influence of Frequency on the Lexical Organization of Verbs

Daniel Cerato Germann[1] **Aline Villavicencio**[2] **Maity Siqueira**[3]
[1]Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[2]Department of Computer Sciences, Bath University (UK)
[3]Institute of Language Studies, Federal University of Rio Grande do Sul (Brazil)
{dcgermann,avillavicencio}@inf.ufrgs.br, maitysiqueira@hotmail.com

Abstract

This work extends the study of Germann et al. (2010) in investigating the lexical organization of verbs. Particularly, we look at the influence of frequency on the process of lexical acquisition and use. We examine data obtained from psycholinguistic action naming tasks performed by children and adults (speakers of Brazilian Portuguese), and analyze some characteristics of the verbs used by each group in terms of similarity of content, using Jaccard's coefficient, and of topology, using graph theory. The experiments suggest that younger children tend to use more frequent verbs than adults to describe events in the world.

1 Introduction

The cognitive influence of frequency has been proven strong in the learning process of both sense and nonsense words (Howes and Solomon, 1951; Solomon and Postman, 1952). Frequency has also been shown to highly correlate with semantic factors, endorsing its importance, through the so called "light verbs" (Goldberg, 1999).

In this study, we investigate whether words that are more frequent have a higher chance of earlier acquisition. For this purpose, we compare data from children and adults, native speakers of Brazilian Portuguese, on an action naming task, looking at lexical evolution, using statistical and topological analysis of the data modeled as graphs. Our approach innovates in the sense that it directly simulates the influence of a linguistic factor over the process of lexical evolution.

This paper is structured as follows. Section 2 describes related work. Section 3 presents the materials and methods employed. Section 4 presents the results and section 5 concludes.

2 Related Work

Steyvers and Tenenbaum (2005), use some properties of language networks to propose a model of semantic growth, which is compatible with the effects of age of acquisition and frequency, in semantic processing tasks. The approach proposed in this paper follows Steyvers and Tenenbaum in the sense of iterative modifications of graphs, but differs in method (we use involutions instead of evolutions) and objective: modifications are motivated by the study of frequency instead of production of a topological arrangement. It also follows Deyne and Storms (2008), in directly relating linguistic factors and graph theory metrics, and Coronges et al. (2007), in comparing networks of different populations.

This study also follows Tonietto et al. (2008) in using data from a psycholinguistic action naming task. However, the analysis is done in terms of graph manipulation, instead of pure statistics.

3 Materials and Methods

3.1 The Data

The action naming task was performed by different age groups: 55 children and 55 young adults. Children's data are longitudinal; participants of the first data collection (G1) aged between 2;0 and 3;11 (average 3;1), and in the second collection (G2), between 4;1 and 6;6 (average 5;5) as described by Tonietto et al. (2008). The adult group is unrelated to the children, and aged between 17;0 and 34;0 (average 21;8). Participants were shown 17 actions of destruction or division (Tonietto et al, 2008) and asked to describe it.

Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL 2010, pages 19–23,
Uppsala, Sweden, 16 July 2010.

Data processing and justification of the chosen domain are described in Germann et al. (2010).

The answers given by each participant were collected and annotated with two frequency scores, each calculated from a different source. The first, Fscore, is the number of occurrences of the verb in the "Florianópolis" corpus (Scliar-Cabral, 1993; MacWhinney, 2000). The second, Yscore, is the number of given results searching for the infinitive form of the verb in the "Yahoo!" Searcher (http://br.yahoo.com). In the advanced settings, "Brazil" was selected as country and "Portuguese" as language. Information about these two scores for each group is shown in Table 1.

	G1	G2	G3
Average type Fscore	44.05	35.92	17.84
Average token Fscore	43.44	35.71	21.22
Average type Yscore	15441904	18443193	10419263
Average token Yscore	10788194	9277047	8927866

Table 1: Type and token scores[1].

All scores but type Yscore, decrease as age increases, which is compatible with the hypothesis investigated.

3.2 Simulation Dynamics

Linguistic production of each group was expressed in terms of graphs, whose nodes represent the mentioned verbs. All verbs uttered for the same video were assumed share semantic information, and then linked together, forming a (clique) subgraph. The subgraphs were then connected in a merging step, through the words uttered for more than one video.

To investigate the influence of frequency on the language acquisition process, we used it to change the network over time. Network involution, the strategy adopted, works in the opposite way than network growth (Albert and Barabási, 2002). Instead of adding nodes, it takes an older group graph as the source and decides on the nodes to iteratively remove (taking the younger group graph only as a reference for comparison).

Verbs were ranked in increasing order of frequency. At each step of graph involution, the less frequent verb was selected to be removed, and the resulting graph was measured. Results are reported in terms of the averages of 10-fold cross-validation (because ties imply in random selection).

Graph theory metrics were used to measure structural similarity: average minimal path length (L), density (D), average node connectivity (k) and average clustering coefficient (C/s)[2]. In the involution, k and D, measure semantic share, since that is what relations among nodes are supposed to mean (see above). L and C/s are intended to measure vocabulary uniformity, since greater distances and lower clusterization are related to the presence of subcenters of meaning.

In order to compare the contents of each graph as well, we employed a measure of set similarity: Jaccard's similarity coefficient (Jaccard, 1901). Given two sets A and B, the Jaccard's coefficient J can be calculated as follows:

$$J(A, B) = \frac{x}{(x+y+z)},$$

where "x" is the number of elements in both A and B, "y" is the number of elements only in A, and "z" is the number of elements only in B.

4 Simulation Results

As we remove the verbs with lower frequency from the graph of an older group, the overall structure should approximate to that of a younger group, and both should get more similar concerning content. Therefore, the most relevant part of each chart is the begging: the first removed verbs are expected to be those that differentiate graphs.

4.1 Network Involution Topology

The graph theory metrics are shown in Figures 1 and 2 in terms of 2 lines: network involution (a) by using the selected criterion, and (b) by using random selection (10-fold cross validation). In addition, each figure also shows the measure for the younger group as reference (a dashed, straight, thick line).

In Figure 1, columns represent a graph theory metric, and rows represent the use of a different score. Each legend refers to all charts.

The results for the simulations from G2 to G1, (Figure 1) show that the four metrics are clearly distinct from random elimination from the beginning, indicating that frequency plays a role in the process. C/s is particularly distinct from ran-

[1] Given the measure magnitude, values of Yscore were presented without the decimal fraction.

[2] We adopted the local clustering coefficient of Watts and Strogatz (1998), but as the graphs may become disconnected during network modification, this value is further divided by the number of disconnected subgraphs.

dom: while the former remains constant almost to the end, indicating a highly structured (clustered) graph, the later shows effects of graph partitioning. The remaining metrics presented their greatest approximations to the reference line before the middle of the chart, suggesting that the initial verbs were actually the ones differentiating both graphs. These results suggest an initial increase in semantic share, as *k* and *D* increase, and in uniformity, as nodes get closer to one another (*L*) and remain clustered (*C/s*). In Figure 2, the same tendencies are maintained, although not as clearly as the previous results. The greatest approximations of *k* and *D* happen in the first half of the chart, but in a smoother way. *C/s* still behaves steadily, remaining stable during most of the simulation. Yscore resembles Fscore (the same way as in Figure 1), and was not presented due to space restrictions.

4.2 Network Involution Set Similarity

In the Jaccard's coefficient charts, a rise or stabilization means that "different verbs" (present only in the older graph) were eliminated (increasing set similarity), and a descent means that "common verbs" (present in both graphs) were eliminated instead.

Charts for "excluded different" and "excluded common" verbs (and their random counterparts) are presented in percentage. By doing so, it is possible to measure the exact evolution of both, despite the proportion between them (there are much more "common" than "different" verbs). A rise in the "Excluded Different" line means that sets are getting similar, while stabilization (descents are not possible) means that they are getting different. The opposite applies to the "Excluded Common" line.

In the figures, charts are arranged in columns (the score being used) and rows (the parameter being measured). Each legend is particular to each row (one to Jaccard's coefficient and another to the excluded verbs).

Both simulation sets (Figures 3 and 4) confirm the expected pattern in general: an initial increase in the proportion between "different" and "common" verbs. In Figure 3, Yscore presents an unexpected descent just before the middle, followed by a sharp rise. Since the greatest descent happens just in the end, we interpret this middle descent as data noise. In Figure 4, Fscore presents an almost random result, indicating that the score had low impact in content similarity for this simulation. Fscore in Figure 3 and Yscore in Figure 4 behaved as expected, with most "differ-

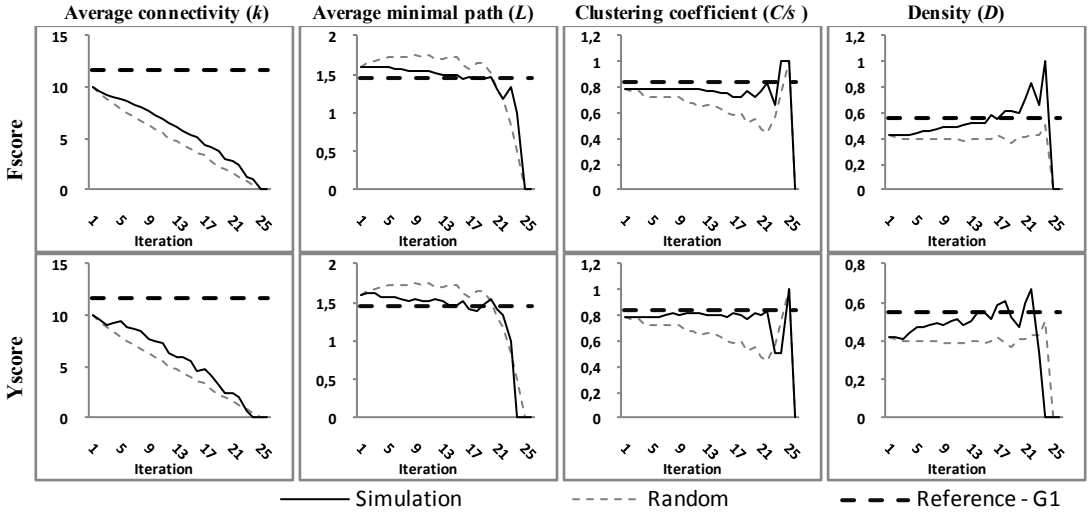

Figure 1. Involution from G2 to G1 using three scores for node removal: graph theory metrics.

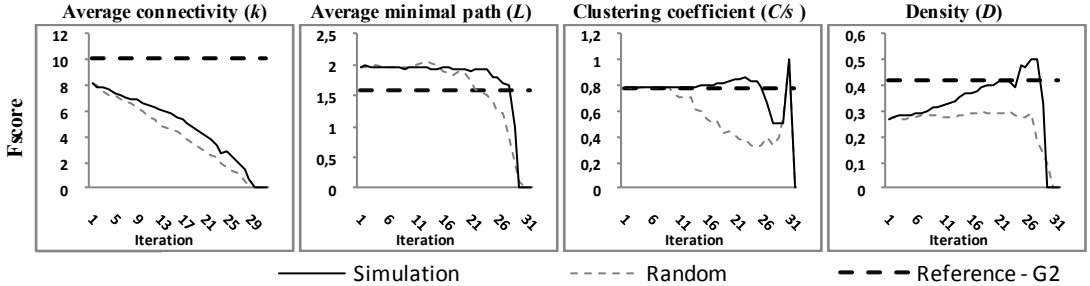

Figure 2. Involution from G3 to G2 using three scores for node removal: graph theory metrics.

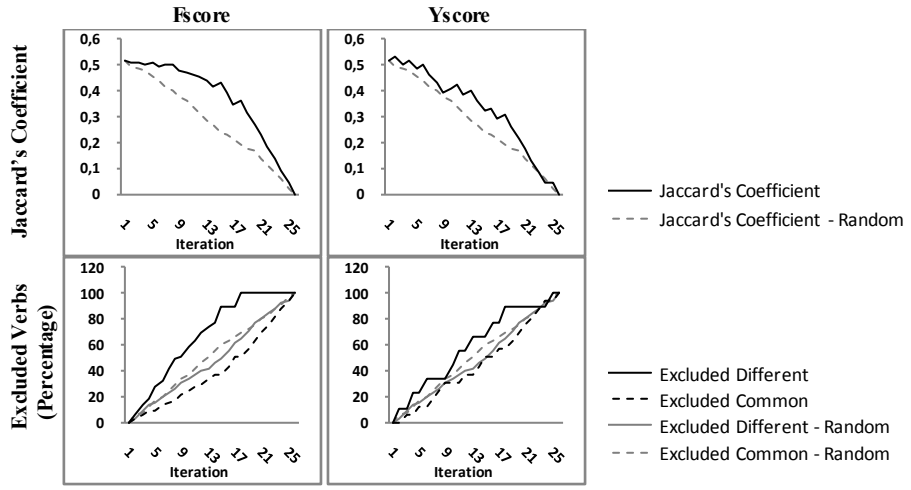

Figure 3. Involution from G2 to G1 using three scores for node removal: set theory metrics.

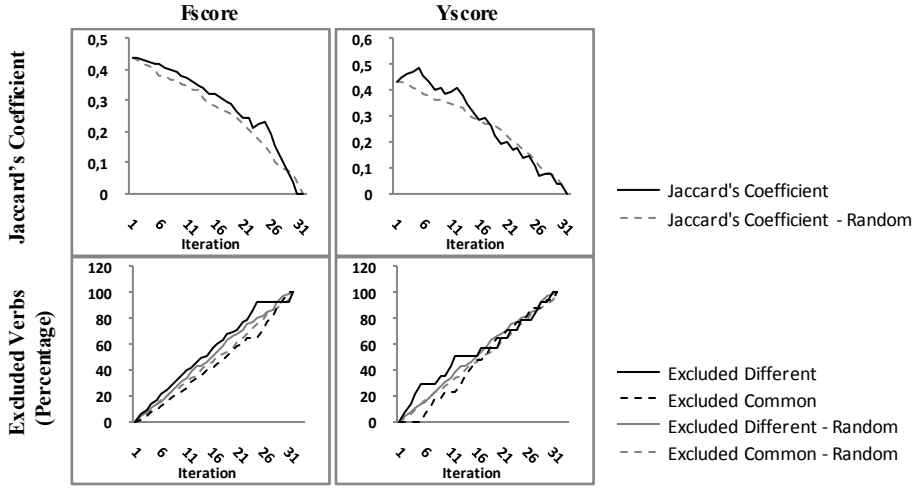

Figure 4. Involution from G3 to G2 using three scores for node removal: set theory metrics.

ent" verbs being excluded before the middle of the chart. Jaccard's coefficient follows the same pattern.

5 Conclusions and Future Work

This study has investigated the influence of frequency on verb acquisition and organization using both graph and set theory metrics. In general, results from the topological analysis showed a tendency towards the reference value, and the greatest similarities were mostly collected in the beginning, pointing for a preference of children to use verbs more frequently perceived in the language. So we conclude that both the model of involution and the given analysis are appropriate for linguistic studies concerning vocabulary evolution[3].

For future work, we intend to apply the same approach to other parameters, such as concreteness, and syntactic complexity (and combinations, and to investigate lexical dissolution in the context of pathologies, such as Alzheimer's disease, and in larger data sets, in order to further confirm the results obtained so far.

Acknowledgments
This research was partly supported by CNPq (Projects 479824/2009-6 and 309569/2009-5), FINEP and SEBRAE (COMUNICA project FINEP/SEBRAE 1194/07). We would also like to thank Maria Alice Parente, Lauren Tonietto, Bruno Menegola and Gustavo Valdez for providing the data.

[3] Since the measures were taken from the whole graph, it is not possible to determine a measure of significance. However, the comparisons with random elimination can be seen as a tendency. Additionally, the experiments consist of two simulations, over three different data sets, using two different sets of frequency (and a combination with polysemy) and two kinds of metrics, which provide robustness to the results.

References

Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47-97.

Kathryn A. Coronges, Alan W. Stacy and Thomas W. Valente. 2007. Structural Comparison of Cognitive Associative Networks in Two Populations. *Journal of Applied Social Psychology*, 37(9): 2097-2129.

Simon de Deyne and Gert Storms. 2008. Word associations: Network and semantic properties. *Behavior Research Methods*, 40(1): 213-231.

Daniel Cerato Germann, AlineVillavicencio and Maity Siqueira. In press. An Investigation on Polysemy and Lexical Organization of Verbs. In *Proceedings of the NAALHT - Workshop on Computacional Linguistics 2010*.

Adele E. Goldberg. The Emergence of the Semantics of Argument Structure Constructions. 1999. In *Emergence of Language*. Lawrence Erlbaum Associates , Mahwah, NJ.

Davis H. Howes and Richard L. Solomon. 1952. Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology*, 41(6): 401-410.

Paul Jaccard. 1901. Distribution de la flore alpine dans le Bassin des Drouces et dans quelques regions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles,* 37(140): 241–272.

B. MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ.

Scliar-Cabral . 1993. *Corpus Florianópolis*. Retrieved January 10, 2009, from http://childes.psy.cmu.edu/data/Romance/Portuguese/Florianopolis.zip

Richard L. Solomon and Leo Postman. 1952. Frequency of usage as a determinant of recognition thresholds for words. *Journal of Experimental Psychology*, 43(3): 195-201.

Mark Steyvers and Joshua B. Tenenbaum. 2005. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science: A Multidisciplinary Journal*, 29(1): 41-78.

Lauren Tonietto et al. 2008. A especificidade semântica como fator determinante na aquisição de verbos. *Psico*, 39(3): 343-351.

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