UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

ANDERSON ROBERTO SANTOS DOS SANTOS

A Computational Investigation of Verbs during aging with and without Alzheimer's Disease

Thesis presented in partial fulfillment of the requirements for the Master's degree in Computer Science

Profa. Dra. Aline Villavicencio
Advisor

Profa. Dra. Jerusa Fumagalli de Salles
Co-advisor

Porto Alegre
2011

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Santos, Anderson R. S. dos

A Computational Investigation of Verbs during aging with and without Alzheimer's Disease / Anderson Roberto Santos dos Santos. – 2011.

67 p.:il.

Orientador: Aline Villavicencio; Co-orientador: Jerusa Fumagalli de Salles.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2011.

1. Natural language processing. 2. Cognitively based models. 3. Mental lexicon. I. Villavicencio, Aline. II. Jerusa, Salles III. Título.

ACKNOWLEDGEMENTS

This work would not be possible without some help. Therefore, I would like to thank:

To CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for sponsoring this work;
To my parents, Roberto Couto and Maria Conceição, for their support and tenderness;
To my advisor, Aline Villavicencio – who trusted me since the beginning – for her wise advice, her friendship and her comprehension;
To Jerusa Salles, my co-advisor, for her valuable psycholinguistic bibliographic contribution;
To Marco Idiart for good insights during the work;
To Maria Alice Parente for providing the data for this research;
To Ronnie Alves, for his help with graph tools;
To the PPGC staff, for their patience and their kind help on bureaucratic issues;
To my friends and lab co-workers (Rodrigo Wilkens, Mario Machado, Gustavo Valdez and Paulo Schreiner), always there, giving some insights and helping;
To Breno Fernandes, Fernando Mariano, Hadson Batista, José Cabaleiro, Luiz Tanajura, Matheus Quadros, Rafaele Mattos, Roberta Pinheiro and Tássio Coêlho, always present, in the good and the difficult moments, with wise words, good ideas, loyal friendship and teaching me the happiness of the simple things;
To my ex-girlfriend, Fernanda Bataiolli, for her love, understanding, and sweet presence;
To all my other friends, the family that I chose to have;
To CNPq(Projects 479824/2009-6 and 309569/2009-5), FINEP, SEBRAE (1194/07) and FAPERGS (08513752) for partially supporting this research.

ABSTRACT

Alzheimer’s disease produces alterations of cognitive functions and of processes that are responsible for language and memory. In order to have a better understanding of language changes, we investigate the characteristics of the semantic networks of patients diagnosed with probable Alzheimer, focusing on verbs. The results of comparisons with networks of healthy individuals and patients with Alzheimer disease highlight some topological differences among them.

We also constructed classifiers that could capture the differences between the various profiles of speakers, and that can be used to classify unknown speakers according to the closest profile. We made this effort in order to help the diagnosis of diseases that affect language, such as the Alzheimer’s disease.

Keywords: natural language processing, cognitively based models, mental lexicon, decline of the verbal lexicon, Alzheimer.

Uma Investigação Computacional do uso de verbos no envelhecimento com e sem doença de Alzheimer

RESUMO

A doença de Alzheimer produz alterações nas funções cognitivas, entre eles, de processos que são responsáveis pela linguagem e memória. Com o intuito de termos uma melhor compreensão das alterações da linguagem, este trabalho investigou características presentes em redes semânticas de pacientes com diagnóstico de provável Alzheimer, com foco nos verbos. Os resultados das comparações entre as redes de indivíduos saudáveis e pacientes com Alzheimer indicam diferenças topológicas entre eles.

Neste trabalho, também foram construídos classificadores que poderiam captar as diferenças entre os vários perfis de indivíduos, e que podem ser utilizados para classificar novos indivíduos de acordo com o perfil mais próximo. Esse esforço se deu com o intuito de ajudar no diagnóstico de doenças que afetam a linguagem, como a doença de Alzheimer.

Palavras-Chave: processamento de linguagem natural, modelos cognitivamente motivados, léxico mental, declínio do léxico verbal, Alzheimer.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AD	Alzheimer’s disease
ADG	Alzheimer’s Disease Group
AI	Artificial Intelligence
DRS	Disability rating scale for severe head trauma patients
ENC	Elderly Normal Controls
GA	Genetic Algorithm
HEG	Healthy Elderly Group
HYAG	Healthy Young Adult Group
HYP	Hyponymy/Hypernymy
IDE	Integrated Development Environment
LMT	Logistic Model Trees
MER	Meronymy/Holonymy
MNC	Middle Aged Controls
NC	Normal Control
NLP	Natural Language Processing
POL	Polysemy
SVM	Support Vector Machines
WWW	World Wide Web

TABLE OF CONTENTS

INTRODUCTION

Language is one of the features that distinguishes humans from animals (HAUSER *et al.*, 2002). And even though there are thousands of words and unlimited ways of joining them together, the mind can deal with language so efficiently that current Natural Language Processing (NLP) technology has not yet been able to replicate it computationally. (KE, 2007). NLP is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages (CHARNIAK; MCDERMOTT, 1985).

An adequate modeling of language requires understanding of the theories of cognitive processes of its learning and loss and their application in a computational investigation of cognitive theories can lead to the confirmation or refusal of characteristics of these theories and to more robust theories and models of language. In addition, the particularities of the context in which these processes occur can influence their development and outcomes. For instance, the cognitively motivated process of language loss can be simply a natural consequence of aging or be triggered by a disease.

A better understanding of language loss is even more crucial due to the changes on the population pyramid after the baby boom decades ago and a higher life expectancy which lead to the need to prevent, diagnose and treat causes of cognitive impairment present in normal and impaired aging (MANSUR *et al.*, 2005).

Alzheimer's disease (AD) is one of the causes of an impaired aging, and it is estimated that 35.6 million people currently suffer from the disease and in 20 years this number will reach 65.7 million of individuals[1], with an estimated overall treatment cost of 315 billion dollars per year in the world. Alzheimer's disease is responsible for more than 50% of the cases of dementia, and it is one of the pathologies that cause, among other consequences, alteration of cognitive functions and of the processes responsible for language and memory (MANSUR *et al.*, 2005).

In relation to language capacity, previous studies have found a progressive deterioration in performance in phonetic-phonological, syntactic, semantic and pragmatic-discursive processes in the context of Alzheimer's disease (MAC-KAY *et al.*,

[1] Figures from the Alzheimer's Disease International, 2009.

2003; MANSUR *et al.*, 2005; ORTIZ, 2009). For instance, Alzheimer's disease causes, among other consequences:

- memory impairment (impaired ability to learn new information or recall previously learned information);
- aphasia (inability to use or understand language because of a brain lesion);
- apraxia (impaired ability to perform motor activities despite intact motor function) ;
- agnosia (inability to recognize or identify objects despite intact sensory function) (MATTIS, 1976);

The disease also causes the disturbance in executive functioning (i.e., planning, organizing, sequencing, abstracting) (MATTIS, 1976).

In the context of semantic memory, there is no consensus about the precise nature of the changes in AD (MANSUR *et al.*, 2005). Based on the results of semantic memory tests such as the Hodges Battery (HODGES *et al.*, 1992; HOWARD; PATTERSON, 1992), two main theories are proposed to explain the semantic deficits of cognitive performance on these explicit semantic tests. The first one proposes a degradation of the semantic memory itself, while the second advocates for a failure to retrieve information from memory (MANSUR *et al.*, 2005; ROGERS; FRIEDMAN, 2008).

There seems to be a preference for more general and frequent verbs (BREEDIN et al., 1998; THOMPSON, 2003; KIM; THOMPSON, 2004; BARDE et al., 2006; THOMPSON; SHAPIRO, 2007), which may be due to the fact that these verbs are applicable in many distinct situations.

Closely related factors such as polysemy and synonymy are also seen as having an important role in the human learning process (HILLS *et al.*, 2009). Features like these may influence the organization of the mental lexicon, e.g., from the need of fast concept retrieval (STEYVERS; TENENBAUM, 2005).

On the one hand, in studies using graph analysis of language in general, there are evidences of the presence of some common features on the knowledge organization of the mental lexicon (STEYVERS; TENENBAUM, 2005). On the other hand, other works, also using graph analysis to model semantic knowledge of AD patients, have found convincing evidences that the network structure is affected, such as having more unnecessary connections, and concepts being organized in a relatively chaotic way (Chan, Butters, Paulsen, et al., 1993; Chan, Butters, & Salmon, 1997; Chan, Butters, et al., 1995;

Chan, Butters, Salmon, et al., 1993; Chan, Salmon, et al., 1995; Chan, Salmon, Nordin, Murphy, & Razani, 1998).

In order to model the semantic knowledge as graphs, Chan, Butters, & Salmon (1997) built an individual semantic network using a Pathfinder (DEARHOLT; SCHVANEVELDT, 1990) analysis based on subjects' responses to a triadic comparison task with twelve animals. The triadic comparison task consists in showing three animals at a time and the individual is asked to point the two animals that are most alike. The comparison task generates proximity data values between the animals. These proximity data were used in a Pathfinder analysis to create the individual networks. Subsequently, a similarity index was found based on the average Closeness[2] measure (GOLDSMITH; DAVENPORT, 1990) of each AD patient's semantic network and the standard healthy elders' network (Chan, Salmon, et al., 1995).

In this work we investigate the characteristics of semantic networks of AD patients[3], focusing on the lexical organization of verbs. The hypothesis that we investigate in this work is whether it is possible to find changes in the global structure of semantic networks that reflect differences among distinct groups of people; in particular, elders with and without Alzheimer's disease.

We use psycholinguistic data from an action naming task, comparing the output of AD patients with those from healthy individuals. We represent the data as semantic networks, which seem to play an important role in the modeling of the organization of lexical knowledge and have been used to describe access to the mental lexicon (STEYVERS; TENENBAUM, 2005). We analyze collective[4] semantic networks using statistical and topological measures present in complex network theories.

[2] The Closeness measure in (GOLDSMITH; DAVENPORT, 1990) is based on the Jaccard similarity coefficient of the edges of the node and is not a centrality measure used in graph theory.

[3] Due to the impossibility of detecting the presence of histological brain features in living elderly individuals, the diagnosis is of probable or possible Alzheimer Disease (McKhann et al., 1984).

[4] Collective networks are modeled using a group of individuals, rather than only one.

This work also aims to investigate if complex network measures are a good approach for pathology investigation. Could the structural characteristics of semantic networks of verbs in terms of graph statistical features be used to extract useful information to predict the presence of Alzheimer disease in native speakers? Could these measures be used to elucidate questions about language impairments caused by pathologies?

This work attempts to advance towards a better integration of findings in cognitive linguistic work and NLP developments, for the construction of more adaptive and cognitive-based NLP technology on one side, and the empirical testing of linguistic theories through computational simulations on the other. For that, we propose the implementation of a collective semantic network impact model, based on a new comparison methodology, which can be used to help the diagnosis of AD. This can be achieved by training machine learning classifiers, using the data from an action naming task to predict membership of a speaker's belonging to a healthy or AD group.

This dissertation is organized as follows: chapter 2 reviews the theoretical issues underlying the work, as well as some related works. Chapter 3 shows how the psycholinguistic data are acquired and modeled as collective networks; it also presents the data collection software created in order to collect new data, and discusses the semantic network analysis software created in this work, presenting its motivation and architecture. Chapter 4 presents the experiments done using the results from the psycholinguistic tasks, and results obtained, addressing questions raised along the work. Finally, in chapter 5, we discuss the conclusions and future works.

RELATED WORKS

In this chapter the theoretical background used in this work is discussed in detail. First we give an overview of semantic networks, the formalism used to represent psycholinguistic data in this work. Then we present some complex network measures and their applications to semantic networks. Subsequently, we look at some related work on Alzheimer's disease, focusing on those employing semantic networks.

1.1 Complex and Semantic Networks

Networks are structures present in many life systems: from neural systems to food webs (NEWMAN, 2004), and we live in a world full of them (KE, 2007). Complex networks is connected to the study of graphs as the representation of other systems elements (KE, 2007). The links, however, depend on the characteristic that we want to study and reflect intrinsic properties of the modeled system; for example, people may be connected by links of friendship, while cities are connected if they have routes that connect them (KE, 2007). When complex network graphs express relations among concepts, they are often called semantic networks.

There are few main classes of complex networks: random, scale free and small-world networks (STEYVERS; TENENBAUM, 2005). One of the most fundamental network is the random network, and it can be obtained by starting with a set of *n* vertices, and two random vertices have a *p* probability of being connected (ERDÖS, 1959).

After the discovery of some interesting properties of random networks (ERDÖS, 1959), attention to their analysis has risen, stimulated by their presence in many fields and powerful computational tools availability (KE, 2007).

Small-world networks have two main characteristics. One is an average short path length, also found in random networks, which means that, no matter how large the network is, the path between any two nodes in the network has a small number of intermediate nodes (WATTS; STROGATZ, 1998). The second characteristic is a high clustering coefficient (that will be explained later), which is not found in random networks (WATTS; STROGATZ, 1998). The authors also found that the small-world structure is present in several networks from different origins, such as semantic networks, social networks, and World Wide Web (WWW) (BARABÁSI; ALBERT, 1999).

In most cases, a node with a high number of connections, will serve as a shortcut for paths between the nodes and, therefore, will be called hubs (BARABÁSI, 2002).

Scale-free networks are characterized as having a power-law degree distribution, in which there are some relatively few nodes, with many connections, serving as hubs, and most of the nodes with few connections (BARABÁSI; ALBERT, 1999). Therefore, the scale-free networks have a node connectivity distribution (it will be explained later) following a power law (BARABÁSI; ALBERT, 1999).

In psycholinguistic terms, quantitative models of associative networks have been used to explain several priming and interference phenomena in the context of human learning and memory research (DEESE, 1965; COLLINS; LOFTUS, 1975; NELSON et al., 1998; ANDERSON, 2000). In this case priming refers to the implicit memory effect in which exposure to a (perceptual, semantic, or conceptual) stimulus influences response to a subsequent stimulus, such as showing a list of words that includes the word table to a person, and asking her later to complete a word starting with tab, when the probability that the answer is table is greater with priming (KOLB; WHISHAW, 2003)[5].

For instance, Nelson et al (1998) used associative networks to predict the performance of some memory retrieval tasks. As a result of these works, a few general characteristics have been discovered, such as some of the processes that involve the search and formation of semantic memories (ANDERSON, 2000).

The semantic meaning is not separated from the structure of the network, and some statistical measures which reflect the structural principles, can reveal the nature of the semantic inside the network (STEYVERS; TENENBAUM, 2005).

Recent works reported a small number of global structural characteristics that are present in language networks (as well as in several biological networks) (SOLÉ; FERRER I CANCHO, 2001; DOROGOVTSEV; MENDES, 2002; MOTTER et al., 2002; SIGMAN; CECCHI, 2002; STEYVERS; TENENBAUM, 2005; SOLE et al., 2006; KE,

[5] Another example is when people see an incomplete sketch that they are unable to identify, and are then shown more of the sketch until they recognize the picture. Later, they will identify the sketch at an earlier stage than it was possible for them before (KOLB; WHISHAW, 2003).

2007). Some of the features reported in these works are small-world[6] (WATTS; STROGATZ, 1998) and scale-free (BARABÁSI; ALBERT, 1999).

1.1.1 Basic Concepts of Graph Theory

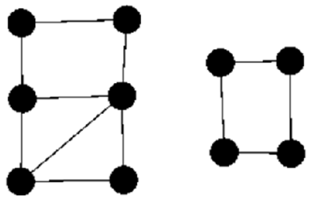
We need to introduce some basic concepts of graph theory in order to explore complex networks and the studies in the field.

Modeled as a graph, a semantic network is composed of nodes that can be mapped to concepts. Links between the nodes indicate some relation between them (QUILLIAN, 1968). For instance, a synonym dictionary can be transformed into a network by modeling words as nodes, and if two words are synonyms they have a link in the network. These networks can be directed or undirected. In directed graphs, the link between two nodes is an arc expressing the direction of the relation. A directed network has an undirected network counterpart when the arcs are replaced by edges (STEYVERS; TENENBAUM, 2005).

Two nodes connected by a link are neighbors. A path is a sequence of edges that connect one node to another. We will refer to them as the distance of two nodes A and B by the shortest path between them. In real networks there is not always a path between two nodes. However, for the statistical analysis, the biggest connected component (STEYVERS; TENENBAUM, 2005) is used. A connected component is a graph (or part of it) where there is always a path for any two nodes within it. Figure 0.1shows a network with two connected components.

[6] Complex networks have been gaining momentum recently. The key of the research lies in the simple explanations that can describe the structure and dynamics of seemingly complex real-life networks. The research field turned out to be of significance to fields as diverse as economics and ecology. Watts' book, *Small worlds: The dynamics of networks between order and randomness* (WATTS, 1999) has proven to be one of the starting points of the recent surge of interest in this field. Since then, a number of papers have been published in this area.

Figure 0.1: A network with two connected components.



Fonte: Santos (2011).

There are four important statistical features that can be defined using the previous terminology: the average distance *L*; the diameter *D*; the clustering coefficient *C*; and the degree distribution *P(k)*.

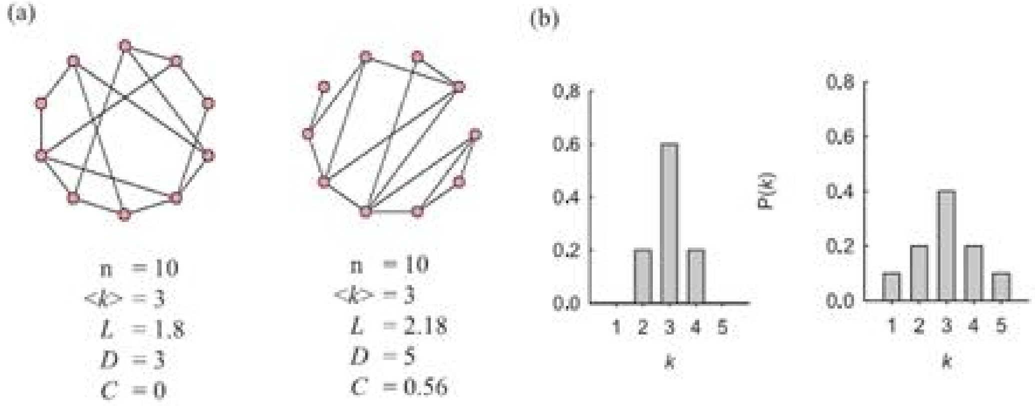
- n is the number of nodes in the network.
- k_i is called as the degree of the node i and expresses the number of links of the node i.
- $<k>$ is the average degree of the network nodes.
- L is the average of the shortest path lengths among all pairs of nodes in the network.
- D (also called as the diameter of the network) is the maximum of these distances among all pairs of nodes, which means that, at most, D steps are required to travel between any two nodes.
- The clustering coefficient *C* is the level of local clustering, and represents the probability that two random nodes are neighbors (STEYVERS; TENENBAUM, 2005). First, the clustering is measured for each node, as can be seen in Equation 1 (Watts & Strogatz, 1998). In the equation 1, T_i expresses the number of existing connections among the neighbors of node i. k_i is the number of neighbors of i. Therefore, $k_i(k_i - 1)/2$ is the maximum possible number of connections among the neighbors of i. One can see that T_i can never exceed $k_i(k_i - 1)/2$, and that the clustering coefficient *C* of the node is normalized between 0 and 1. Even though the clustering coefficient is sensitive to the number of connections, it is possible that two networks with the same number of links have different clustering coefficients, as can be seen in Figure 0.1 (STEYVERS; TENENBAUM, 2005).

$$C_i = {}^{T_i}\Big/_{\binom{k_i}{2}} = {}^{2T_i}\Big/_{k_i(k_i-1)} \qquad (1)$$

- The degree distribution *P(k)* is the probability that a random node will have the degree k (i.e. having k neighbors) (STEYVERS; TENENBAUM, 2005). This feature is exploited in a better way when the full distribution of *P(k)* is plotted as a function of k (STEYVERS; TENENBAUM, 2005). The shape of these plots can show special signatures of different kinds of network (STEYVERS; TENENBAUM, 2005).

Figure 0.1 illustrates some properties of random graphs. First, for a fixed *n* and <*k*> (number of links), high values of *C* tend to imply high values of *L* and *D* (STEYVERS; TENENBAUM, 2005). Second, the degree distribution is approximately bell shaped. These two features are well present in random graphs, but not in some natural networks, such as semantic networks.

Figure 0.1: An illustration of the graph-theoretic properties that are applied to semantic networks (STEYVERS; TENENBAUM, 2005).



Fonte: STEYVERS; TENENBAUM (2005).

In Figure 0.1.a, there are networks with equal numbers of nodes and edges. For both networks, the variables *n* (number of nodes) and <*k*> (average degree) are shown, as well as the statistical properties: *L* (average shortest path length), *D* (diameter) and *C* (clustering coefficient). Note that the two networks have different clustering coefficients, even though they have the same *n* and <*k*>. Figure 0.1.b, are degree distributions corresponding to the two networks in Figure 0.1.a. Both networks show the typical pattern for random graphs: approximately bell-shaped distributions (STEYVERS; TENENBAUM, 2005). The summary of the metrics presented in this section is in Table 0.2.

- Centrality: There are various measures of the centrality of a vertex within a graph that can determine the relative importance of that vertex within the graph. These measures usually use distances based on metrics. Some important centrality metrics that will be used in this work are shown in Table 0.1. The Betweenness and Closeness metrics showed to be capable of finding some network strengths and weaknesses in a power transmission network system (CADINI *et al.*, 2009).

Table 0.1: Some important centrality measures

Metric	**Description**
Betweenness Centrality	Measures how often a node appears in the shortest paths between nodes in the network (FREEMAN, 1979).
Closeness Centrality	The average distance from a given node to all other nodes in the network (FREEMAN, 1979).
Eccentricity	The distance from a given starting node to the farthest node from it in the network (BOUTTIER *et al.*, 2003).

Fonte: Santos (2011).

- PageRank is a method for network analysis giving numerical weights to each node in the network, in order to measure their importance (PAGE *et al.*, 1998). If node A has an edge pointing to node B, then it is considered that node A is referring to node B like a voting system (PAGE *et al.*, 1998). The more votes that are cast for a node, the more important the node must be. Also, the importance of the node that is casting the vote determines how important the vote itself is (PAGE *et al.*, 1998). The algorithm calculates a node's importance from the votes cast for it. How important each vote is taken into account when a node's PageRank is calculated.

Table 0.2: Some terms and definitions used in the work.

Term/variable	**Definitions**
n	Number of nodes
L	The average length of the shortest path between pairs of nodes
D	The diameter of the network

C	The clustering coefficient (see Equation 1)
k	The degree
$P(k)$	The degree distribution
$<k>$	Average degree
γ	Power law exponent for the degree distribution
random graph	Network where each pair of nodes is joined by an edge with probability p
Small-world structure	Network with short average path lengths L and relatively high clustering coefficient C
Scale-free network	Network with a degree distribution that is power-law distributed

Fonte: Santos (2011).

1.1.2 Lexicon Networks

In Steyvers & Tenenbaum (2005), the authors analyzed the large scale structures of three kinds of semantic networks: word associations of naïve subjects (NELSON *et al.*, 1999), WordNet (MILLER, G. A. *et al.*, 1990), and the Roget thesaurus (ROGET, 1911).

In the associative network built from the word associations, two kinds of network were modeled, one with directed links and another one, undirected, with the same links as edges. Two words were linked in the Roget thesaurus if the words shared the same semantic category in the thesaurus. In the WordNet network, the words were connected if they had one of the possible relations in the dictionary: homonymy, hyponymy, antonymy and meronymy.

The authors have shown that the three networks have the features of small-world structure, characterized by the combination of short-average path lengths and a high-clustered neighborhood. We can see the summary of the statistic properties of the networks in Table 0.3. For comparison purposes, the authors created random graphs using the same number of nodes and edges of the correspondent network. The L_{random} and

C_{random} variables express the L and C averages of the random graphs with the same size (STEYVERS; TENENBAUM, 2005).

Table 0.3: Summary statistics for semantic networks in (STEYVERS; TENENBAUM, 2005).

Variable	Associative Network		Thesaurus	
	Undirected	Directed	Roget	WordNet
n	5,018	5,018	29,381	122,005
$<k>$	22.0	12.7	1.7	1.6
L	3.04	4.27	49.6	4.0
D	5	10	10	27
C	.186	.186	.875	.0265
γ	3.01	1.79	3.19	3.11
L_{random}	3.03	4.26	5.43	10.61
C_{random}	4.35E-03	4.35E-03	.613	1.29E-04

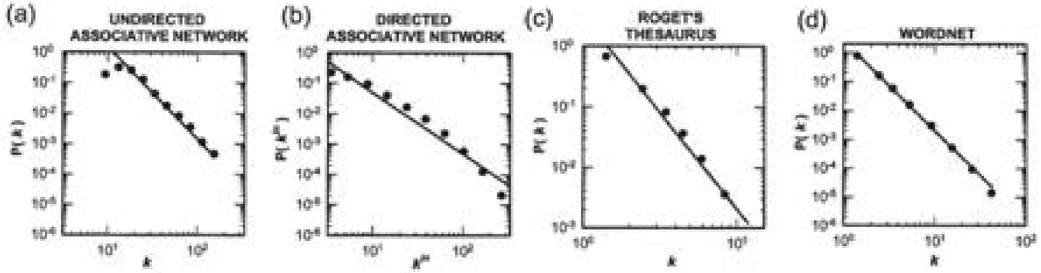
Fonte: STEYVERS; TENENBAUM (2005).

We can see that all the networks are sparse, noticing that a node is connected to a very small percentage of the entire network (around 0.44%), as can be seen by the average degree of the networks ($<k>$) (STEYVERS; TENENBAUM, 2005). Analyzing connectivity, the authors have found that, even being sparse, the networks have almost only one big connected component (STEYVERS; TENENBAUM, 2005). The associative directed network's biggest strongly connected component has 96% of the words; in the same undirected graph, the entire graph is connected. Roget's thesaurus's and WordNet's biggest connected component has approximately 99% of all words (STEYVERS; TENENBAUM, 2005).

When comparing the three undirected networks, we can see that the average shortest-path length (L) grows slower than the size of the network, even though the average number of connections ($<k>$) is lower in bigger networks such as WordNet. The clustering coefficient in these networks has a magnitude order far bigger when compared with random graphs with the same size of edges and nodes (STEYVERS; TENENBAUM, 2005).

The log-log degree distribution plotted in figure 2.3 (from Steyvers & Tenenbaum, (2005)) shows that all networks fit almost perfectly in a power-law function (STEYVERS; TENENBAUM, 2005). The respective exponents of the function were reasonably similar in the three graphs, varying between 3.01 and 3.19 (see Table 0.3), which is consistent with Zipf˝s (1949) findings (STEYVERS; TENENBAUM, 2005). The authors suggest that the small-world and scale-free features present in all language networks are not random and could be related to the human cognitive necessity for the fast retrieval of concepts (STEYVERS; TENENBAUM, 2005). Following the same direction, we analyzed semantic networks of verbs of healthy elders and patients with Alzheimer’s disease in order to explore possible changes in these global features (small-world and scale-free).

Figure 0.2: The degree distributions in the undirected associative network (a), the directed associative network (b), Roget’s Thesaurus (c), and WordNet (d). All distributions are shown in log-log coordinates with the line showing the best fitting power law distribution (STEYVERS; TENENBAUM, 2005).



Fonte: STEYVERS; TENENBAUM (2005).

The work of Motter, de Moura, Lai, & Dasgupta (2002) is also related. They modeled a synonym network using the Moby thesaurus[7] and achieved nearly the same findings as Steyvers & Tenenbaum (2005). We can see their comparison in Table 0.4. Due to the high clustering, and having an associative behavior, human memory tends to keep similar concepts together, allowing an associative search and maximizing efficiency (MOTTER *et al.*, 2002). Table 0.4 also shows the same findings of Steyvers & Tenenbaum (2005) in language networks of several works, sparse networks with a small average minimal path length and high clustering coefficients.

Trying to analyze the relations of nouns in the WordNet database (version 1.6), (SIGMAN; CECCHI, 2002) created a network using some of the combinations of nouns relations:

[7] The dictionary is available at ftp://ibiblio.org/pub/docs/books/gutenberg/etext02/mthes10.zip

- Hyponymy/hypernymy (HYP). Hyponym shares a type-of relationship with its hypernym. For instance, *oak* is a hyponym of *tree*, and *dog* is a hyponym of *animal*. The opposite of a hyponym is a hypernym;
- Meronymy/holonymy (MER). A meronym denotes a constituent part of, or a member of something. That is, X is a meronym of Y if Xs are parts of Ys, or X is a meronym of Y if Xs are members of Ys. For example, 'finger' is a meronym of 'hand' because a finger is part of a hand. Similarly 'wheel' is a meronym of 'automobile';
- Polysemy (POL). It expresses the possible meanings or senses that the word can have.

Four networks were created based on the combination of these three kinds of relations. The authors have found that the network only exhibited the feature of small-world when polysemy was added to it which suggests the importance of polysemy in the construction of mental lexicon (SIGMAN; CECCHI, 2002). In our work, we investigate networks which have links expressing pseudo-polysemy relations (how this pseudo-polysemy relations are obtained will be explained in detail in Chapter 0) between verbs in order to check if the networks of healthy individuals and patients with AD show the same features as the polysemy networks of Sigman & Cecchi (2002).

Table 0.4: A summary of some lexical networks.

Network	n	$<k>$	L	L_{random}	C	C_{random}
(MOTTER *et al.*, 2002)	30,244	60	3.16	2.5	0.53	0.002
(SIGMAN; CECCHI, 2002)	66,025					
HYP			11.9		0.002	$1.2*10^{-4}$
HYP+MER			7.4	10	0.010	$1.2*10^{-4}$
HYP+POL			7	7.6	0.080	$1.2*10^{-4}$
HYP+MER+POL			6	7.2	0.081	$1.2*10^{-4}$
(Ferrer I Cancho, Sole, & Köhler, 2004)						
Czech	33,333	13.4	3.5	4	0.1	$4*10^{-4}$
German	6,789	4.6	3.8	5.7	0.02	$6*10^{-4}$
Romanian	5,563	3.4	3.4	5.2	0.09	$9.2*10^{-4}$

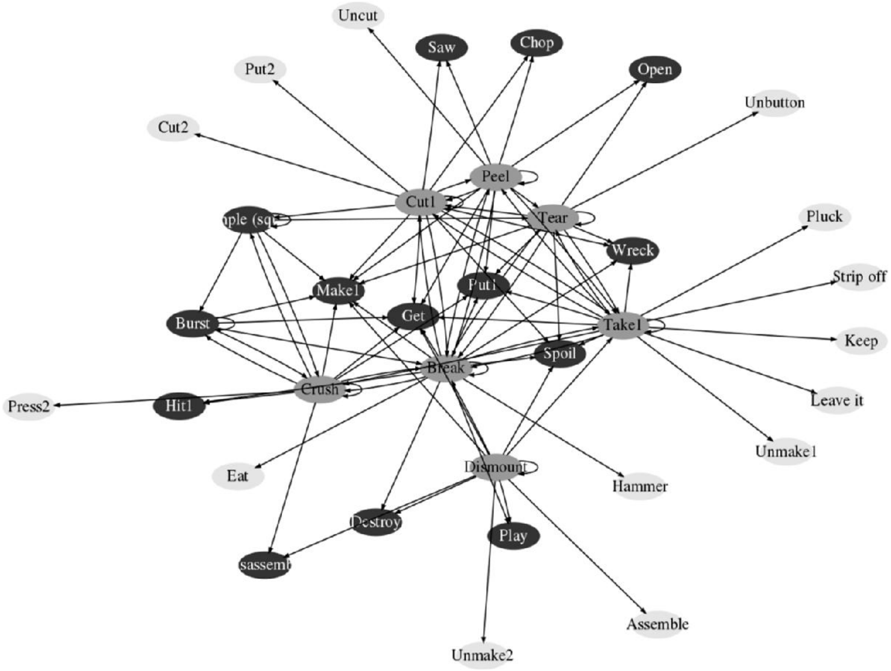
Fonte: Motter, de Moura, Lai, & Dasgupta (2002).

Ferrer I Cancho, Sole, & Köhler (2004) also analyzed semantic networks produced using the thesaurus relations of other languages (Czech, German and Romanian) and confirmed the same findings present in the other works.

Also related to this work is the one of Parente et al. (2001) who made a comparison of semantic networks of adults and children from two languages: Brazilian Portuguese and Mandarin Chinese. The native speakers took part in an action naming task, in which a video with an action performed by an actor was presented (a woman cutting an apple, for instance), and the person was asked to describe what the actor was doing. The test consisted of 17 different actions. The set of verbs of the network is given by all distinct verbs uttered for the videos and if two verbs were uttered for the same video, they were linked together.

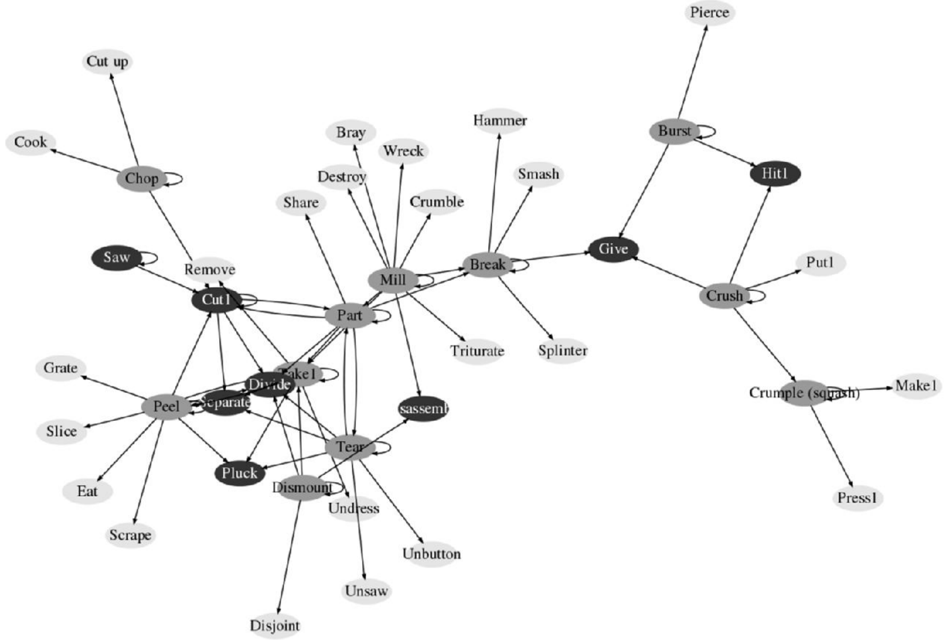
In figure 2.4 and figure 2.5 (from Parente et al. (2011)), we can notice a visible difference in structure between the network of native children and adults of Brazilian Portuguese (PARENTE *et al.*, 2011). That difference reflects on the statistical analysis of these networks. The results obtained indicated that adults preferred to use more specific verbs to each situation and the children often used more common and general verbs (PARENTE *et al.*, 2011). In our work, we follow Parente et al. (2011) using the same action naming task, and the same methodology to create the collective semantic networks.

Figure 0.3: Naming task – Brazilian Portuguese speaking children (PARENTE *et al.*, 2011).



Fonte: Santos (2011).

Figure 0.4: Naming task – Brazilian Portuguese speaking adults (PARENTE *et al.*, 2011).



Fonte: Santos (2011).

One of the cognitive motivations for using networks is that the physical representation of concepts in our neural networks is made by the activation patterns of a concept in a distributed manner, and not by meanings as individual entities (KE, 2007). Instead of only taking into account the low level representation in the brain, we may want to analyze a higher level of semantic representation. And this kind of evaluation can be examined independently (JACOB, 1977).

Moreover, Sigman and Cecchi (2002) claim that polysemy is a consequence of the fast navigation on the mental lexicon. Indeed, Ke (2007) affirms that it could be a consequence of the human cognition (i.e. from metaphoric thinking and generalization) that makes polysemy a consequence of the universal phenomenon of human cognition (LAKOFF, 1987).

As one can see, a better understanding of the universal language concept relations is important to comprehend the human cognition.

1.2 Alzheimer’s

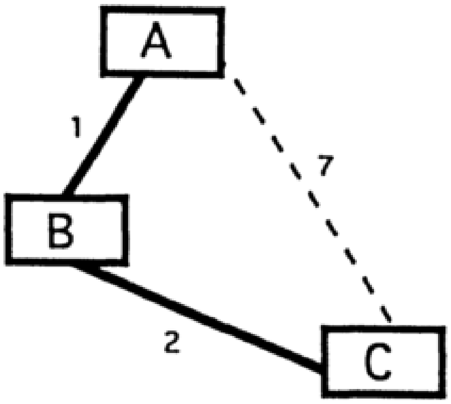
Alzheimer's is a degenerative disease, currently incurable and fatal (BERCHTOLD; COTMAN, 1998). This disease usually affects people over 65 years of age, although younger people might also develop it (ALZHEIMER’S DISEASE INTERNATIONAL, 2009).

Each patient suffers from Alzheimer's disease in a unique way, but there are common points and the most common primary symptom is memory loss (BERCHTOLD; COTMAN, 1998). As the disease advances, new symptoms appear, such as confusion, irritability, aggressiveness, mood changes, language failures, long term memory loss, detachment from reality (BERCHTOLD; COTMAN, 1998; TABERT *et al.*, 2005). Patient die when they start to lose their motor functions (MÖLSÄ et al., 1986). The mean life expectancy after the diagnosis is of approximately seven years (MÖLSÄ et al., 1986).

Although there are several studies about AD in general, there are just a few works in terms of Alzheimer’s semantic network investigation which are of direct relevance to the work proposed in this thesis, (CHAN; BUTTERS; et al., 1995; CHAN; SALMON; et al., 1995; CHAN et al., 1997, 1998). These works share the same data and methodology to create semantic networks. One semantic network is modeled for each individual and is composed of 12 nodes. All nodes represents high-frequency animal names, and are also among the 30 most uttered names in an animal fluency task (dog, cat, cow, horse, rabbit, pig, tiger, lion, bear, elephant, giraffe and zebra) (Chan, Salmon, et al., 1995). Every individual performs a triadic comparison task composed of 220 stimuli, representing all possible combinations of three out of the 12 animal names. For each combination of three animals, the subject is asked to show the two animals that are most alike. The result of the triadic comparison task is a 12x12 matrix containing how many times each pair of animals was indicated as similar.

In order to transform the 12x12 animal comparison matrix in a semantic network, a Pathfinder analysis (DEARHOLT; SCHVANEVELDT, 1990) was used (Chan, Salmon, et al., 1995). The result is a network in which each node represents an animal, and each link represents the distance between the nodes. Two animals that have a high proximity (i.e. are most alike) will be in a short distance from each other, while two animals that have a low proximity will be in a long distance from each other. In the resulting network, two concepts (i.e. animals) will be linked if and only if the distance of their direct link is shorter than the sum of their indirect links (Chan, Salmon, et al., 1995). For example, in figure 2.6 (from Chan, Salmon, et al., (1995)) there is a hypothetical Pathfinder network where the numbers indicate the link weights (i.e. distance) and the dotted lines represent the links that will be removed in the Pathfinder process. A and B, and B and C will not be removed because there is no other path between them with a shorter distance than their direct links. However, there is a path between A and C with a shorter distance (3) than their direct link (7) (DEARHOLT; SCHVANEVELDT, 1990).

Figure 0.5: A hypothetical Pathfinder network for three concepts labeled A, B and C. Concepts A and B, and Concepts B and C, are connected with link lengths of 1 and 2, respectively. Concepts A and C are not directly connected because their link weight (7) is higher than the sum of the indirect links (3) which connect them through Concept B. (Note: solid lines represent direct links.) (Chan, Salmon, et al., 1995).



Fonte: Chan, Salmon, et al. (1995).

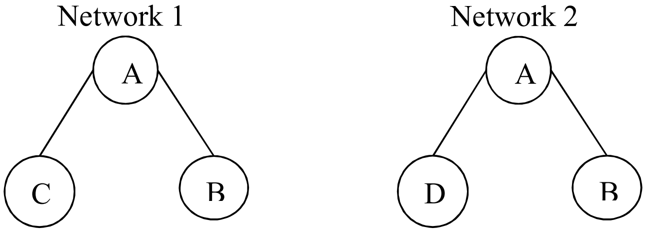
In the Pathfinder network creation process, it is possible to choose the complexity of the resulting network (i.e. number of links) based on two input parameters (DEARHOLT; SCHVANEVELDT, 1990). Parameter *r* indicates the path length, and *q* rules the maximum number of links in the path. Therefore, if we want the simplest network with the minimum number of links, we have to adjust *r* to ∞, and *q* to *n-1* (*n* is the number of concepts that will be represented as nodes) (DEARHOLT; SCHVANEVELDT, 1990).

In the first work using the triadic comparison task and the Pathfinder network methodology (Chan, Salmon, et al., 1995), a network representing the Normal Control (NC) subjects was created using the average matrix values of the NC individuals, and then submitting the matrix to a Pathfinder analysis (adjusting *r* and *q* to produce the minimum number of links); the NC resulting network can be seen in figure 2.8 (from Chan, Salmon, et al. (1995)). A semantic network was created for each AD individual using the individual matrix values and the Pathfinder analysis (Chan, Salmon, et al., 1995). All AD individual networks were compared to the NC network by the calculation of the Closeness measure (GOLDSMITH; DAVENPORT, 1990). This measure, also called Similarity Index, computes the Jaccard similarity coefficient of the node neighborhood for each node. The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets, the formula is Equation (2).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (2)$$

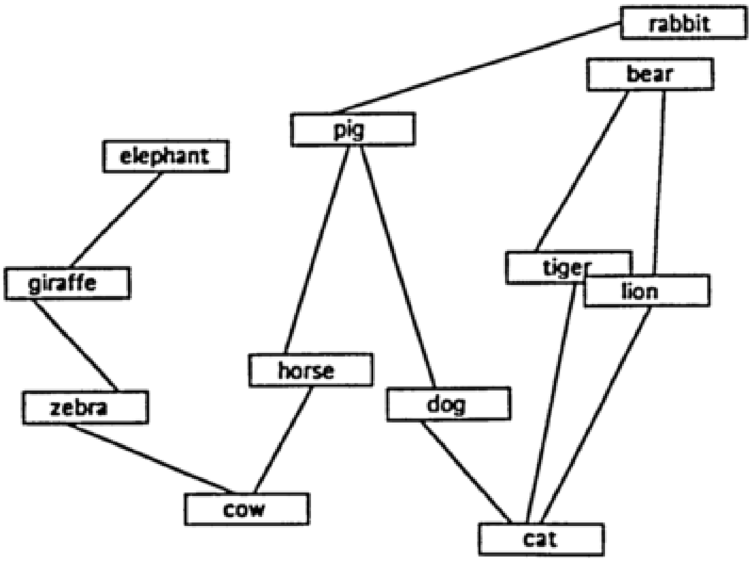
Therefore, the number of common neighborhood concepts is divided by the number of total distinct neighborhood concepts linked to the node in both networks. For example, if the neighborhood concepts of Concept A in network 1 are B and C, and the neighborhood concepts of A in network 2 are B and D (see figure 2.7), then the total number of concepts linked to Concept A is 3 (i.e. B, C and D). However, only Concept B is the common neighborhood of Concept A in both networks. Thus, the Similarity Index of Concept A is 1/3. The Similarity Index between the networks can be calculated by averaging the nodes Similarity Index. The authors (Chan, Salmon, et al., 1995) found that AD Similarity Indices were correlated with the decline in their DRS scores (Disability rating scale for severe head trauma patients) (MATTIS, 1976) over the subsequent year (see figure 2.9 from A. S. Chan, Salmon, et al., (1995)).

Figure 0.6: Networks used as examples to explain how to obtain the similarity index value.



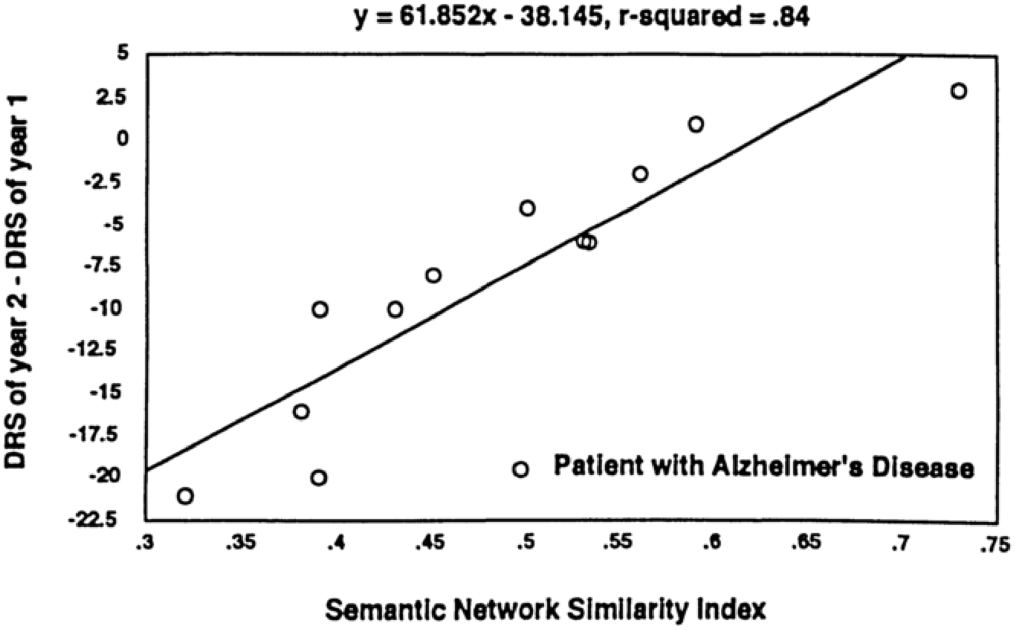
Fonte: Santos (2011).

Figure 0.7: The standard network generated by the Pathfinder analysis of the average proximity data of 12 NC subjects. The derived network is the simplest model that could be developed from this set of proximity data (see text) (Chan, Salmon, et al., 1995).



Fonte: Chan, Salmon, et al., (1995).

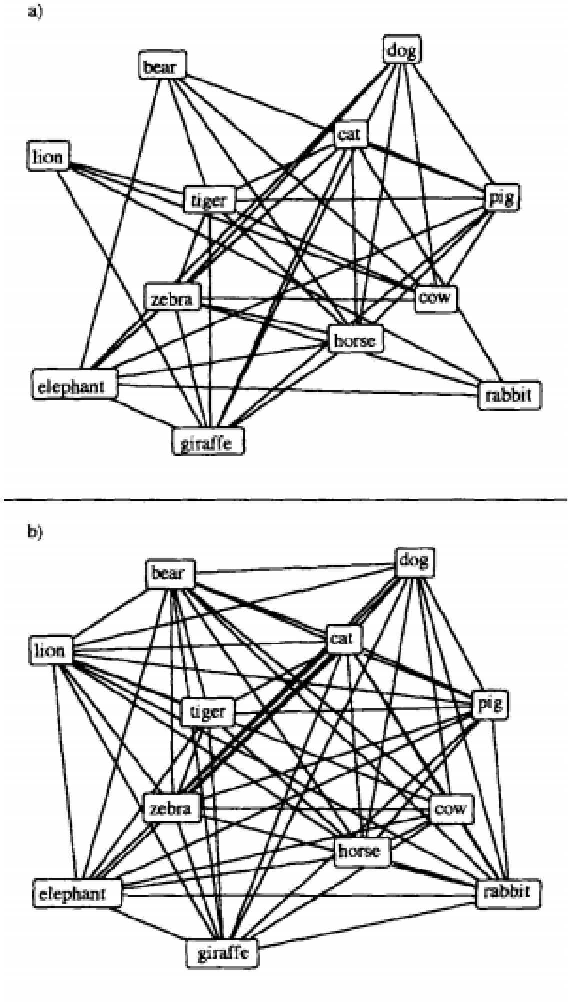
Figure 0.8: The semantic network Similarity Index of 12 AD patients plotted as a function of rate of cognitive decline by the difference between the DRS scores obtained near the time of semantic knowledge testing (year 1) and 1 year later (year 2). The simple linear regression analysis comparing these variables, shown at the top of the Figure, was highly significant (*p* < .001) (Chan, Salmon, et al., 1995).



Fonte: Chan, Salmon, et al., (1995).

Another work analyzing AD semantic networks (Chan, Butters, et al., 1995), built a collective network of AD patients, middle aged controls (MNC), and elderly normal controls (ENC), using the same triad comparison task of the previous work. Then, for each group, an average matrix was derived from the set of individual matrices. Collective networks were generated using a Pathfinder analysis, and setting the input parameters to obtain the most detailed network model (r = 1). The authors found that the weights of each pair of concepts (e.g., cat and dog) in the semantic networks of AD and ENC were not significantly correlated; however, the weights between MNC and ENC were significantly correlated. The AD network has substantially more links than the MNC (respectively 59 and 40 links). This fact is due to the large number of ENC associations with a weight value of 0. As the Pathfinder analysis parameters were settled to obtain the more detailed networks, only the associations with weight 0 were not represented as links. The networks can be seen in figure 2.10 (from Chan, Butters, et al. (1995)). The authors also removed the corresponding links with weight value of 0 between the networks and the result was the same. Thus, the AD networks have more associations, and also atypical strength values in common associations.

Figure 0.9: Semantic networks of (a) 13 elderly normal controls and (b) 13 patients with Alzheimer`s disease generated by Pathfinder analysis (Chan, Butters, et al., 1995).



Fonte: Chan, Butters, et al., (1995).

The number of links in the individual networks is a reflection of their knowledge (SCHVANEVELDT *et al.*, 1985). As individuals possess more knowledge, they develop a more concise network with less unnecessary connections (SCHVANEVELDT *et al.*, 1985). Following the same line of research, Chan, Butters, & Salmon (1997) found that the level of Alzheimer's dementia is significantly correlated with the number of links in AD individual networks. As the severity of the dementia grows, more links are found in the individuals networks (Chan, Butters, & Salmon, 1997).

1.3 General Discussions

There are three main directions in complex network research (NEWMAN, 2003). The analysis of statistical structural properties and their meanings is one of the directions, because new complex measures may lead to new discoveries. The focus on this direction is to find structural properties such as small-world and scale-free, and how to measure and find these properties.

The second direction is the creation of new models of network representation that can lead us to a better understanding of the mental lexicon (NEWMAN, 2003). As discussed by Ke (2007) and Steyvers & Tenenbaum (2005), the links may not be all equally weighted and undirected. Analysis of the force of these connections could help us better understand the relations present in language (KE, 2007) such as polysemy and hyponym. In addition to that, weight values are not static and could change along the learning process (STEYVERS; TENENBAUM, 2005; KE, 2007).

The third approach to networks is to predict the behavior of the systems that generate them, and is more concerned about the structural analysis of these networks. This approach is more interested in how the system is affected by the structural properties.

The modeling of language as complex networks suggests new questions and brings a new perspective to linguists and researchers of related areas, enabling a rich interaction among them. One of the possible exploration fields is the various kinds of aphasias that could be interpreted in a network perspective (KE, 2007). The hypothesis that we investigate in this work is whether it is possible to find changes in the global structure of semantic networks that reflect differences among distinct groups of people; in particular, elders with and without Alzheimer's disease.

In the next sections, the psycholinguistic method used to collect the data will be explained, as well as the model for the generation of semantic networks and the experiments conducted on the data.

THE EXPERIMENTAL SETUP

1.4 Preliminaries

In the previous chapters, the theoretical foundations of this work were discussed, in order to contextualize the theories this work is based upon. We also presented the state of the art of some closely related works. In this chapter, we are going to show how these concepts are adopted in this work.

The experiments share a common methodology: first, the data is collected through psycholinguistic tasks; then, a model is applied to the data to construct and compare semantic networks.

Features will be extracted from the graphs for building a classifier to predict speakers' group membership, which is whether the linguistic production of a new speaker is more similar to that of the group of healthy elders or of that of the AD group. In order to enable this analysis in an integrated environment, a collection of tools were also developed as part of this work.

1.5 Experimental Materials

1.5.1 The Groups

Each of the two groups consisted of 23 individuals - one group with patients diagnosed with probable Alzheimer's disease (AD) and the other one with healthy elders. Prior to the beginning of the study, all participants demonstrated they had the visual, auditory and language abilities required to complete the tasks:

- Alzheimer's Disease Group (ADG): 23 patients[8] diagnosed with probable Alzheimer's disease (Mild AD), with Mean age = 75.6 years; SD = 6.7, and

[8] The size of this sample is compatible with that of other works with Alzheimer's disease: some report from 5 to 11 patients, and others have from 20 to 26 patients (CHAN et al., 1997; BELL et al., 2001; GARRARD et al., 2005; PERAITA et al., 2008; ROGERS; FRIEDMAN, 2008; LAISNEY et al., 2009). This is partly due to the difficulties of finding a larger sample of participants with the same level of the disease (in this case Mild level).

- Healthy Elderly Group (HEG): 23 healthy individuals with Mean age = 72.4 years; SD = 8.2.

In addition, a third group of participants was also considered for evaluation purposes:

- Healthy Young Adult Group (HYAG): with 75 adults (Mean age = 21.69; SD = 3.25)

All participants were native Portuguese speakers and were notified about the objectives and procedures of this research and signed a consent term. The data for sample characterization and for obtaining the inclusion criteria were collected through a Social and Demographic Data Sheet.

1.5.2 The Task of Action Naming

The instrument for the Action Naming Task was developed by Karine Duvignau & Gaume (2004), and consists of 17 short films randomly presented to the participants. Each film lasts between 42 seconds and 1 minute and 13 seconds. In each one of them, a short song plays while a red curtain opens. A woman wearing a clown nose comes from behind the curtain and walks to a table where various objects are placed. She takes one of them and performs an action. The same scenario is presented for 17 different actions as shown in Table 0.1. For more details, see (TONIETTO *et al.*, 2008).

Table 0.1: The 17 actions presented in the films

Film Actions
Peeling a piece of tree trunk
Popping a balloon
Peeling a banana
Peeling a carrot
Taking off the stitches of a shirt’s sleeve
Tearing a newspaper in half
Destroying a Lego® castle
Peeling an orange
Cutting a loaf of bread with a knife
Splitting a loaf with her hands
Crumbling toasted bread
Creasing a piece of paper
Chopping parsley with a kitchen knife

Sawing a board in half
Removing the clothes of a doll
Crushing a tomato with a lid
Breaking a glass with a hammer

Fonte: Santos (2011).

Those actions denoted in the movies were either of destruction or division of an object. The participants were asked “What is the woman doing?”, and the answer given by each participant for each movie was recorded. The answers were manually evaluated and filtered according to the following criteria: all responses that contained verbs were considered valid and maintained if the verb was not unrelated to the main action (e.g., removing “to eat” for the action of sawing a log), and if the answer was not meta-linguistic (e.g. “I don’t know”) or non-verbal (e.g. pointing).

A study done by Tonietto et al. (2008), using the same action naming task presented here, defined degrees of conventionality and specificity for each verb answer given for the movies. The specificity of a verb measures how much the verb can only be used in a few specific cases. In other hand, the conventionality measures how general the verb is in the language. For each verb, in the context of the movie, was established a valid score of specificity, which was previously defined from the judgment of 79 judges, university students of Psychology, Speech Therapy and Education courses. Each student ranked a list of verbs, derived from data collected for the research, in a scale from 1 to 5: (1) highly specific, (2) with tendency towards being specific, (3) average specificity, (4) with trend to be generic and (5) extremely generic. The authors calculated the specificity score for each participant from the average scores for each verb issued. However, the conventionality degree of each verb was found using the same judging methodology but based in the context of each of the movies, and not only in the general context of the verb. The scores were from 1 to 5: (1) highly usual, (2) with tendency towards being usual, (3) average conventionality, (4) with trend to be uncommon and (5) extremely uncommon

To enrich the individual data in this baseline study, the degree of conventionality and specificity defined by (TONIETTO *et al.*, 2008) was added to each answer (verb). We used another frequency degree based on the number of the verb entries found in the *Yahoo*® search engine. The enrichment result example is shown in Table 0.2.

Table 0.2: The individual data used to the baseline classifying experiment

Individual	Move1- Answer	Movie1- Specificity	Movie1- Conventionality	Movie1- Yahoo	Movie2- answer	…
1	Run	1.83	2.46	1670000	Cut	…
2	Walk	4.76	3.83	432000	Kill	…
3	Run	1.83	2.63	1670000	Null	…
4	Go	4.76	3.83	4362000	Make	…

Fonte: Santos (2011).

1.6 Network Modeling

Since we are interested in determining what are the usual characteristics of the language used by a group of speakers, all networks used in this work are collective networks; therefore they were created using a group of individuals. They are also produced using the same modeling methodology described in this section. The input is the linguistic sample of a group of individuals and the output is the collective network.

For each participant the linguistic sample contains his/her answer for each of the movies. This sample is manually pre-processed by a linguist, removing the main verb from each valid answer. An example of pre-processed data is shown in Table 0.3 containing a toy example, where for each individual the columns contain the main verb uttered for each of the movies. The answer can be null if it was not considered valid (using the rules described in Section 1.5.2).

Table 0.3: Toy data sample used to create a network.

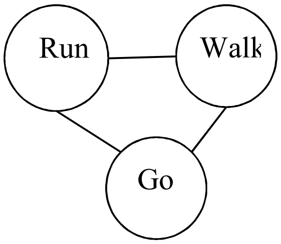
Individual	**Movie-1**	**Movie-2**	**Movie-3**
1	Run	Cut	Jump
2	Walk	Kill	Jump
3	Run	Null	Go
4	Go	Make	Make

Fonte: Santos (2011).

The first step of the network modeling is the node creation. A node is created for each distinct answer uttered for the movie. We consider that verbs spoken for the same movie have some semantics in common, so we make a link among verbs uttered for each movie. The result of this step for “Movie-1” is shown in figure 3.1.

.

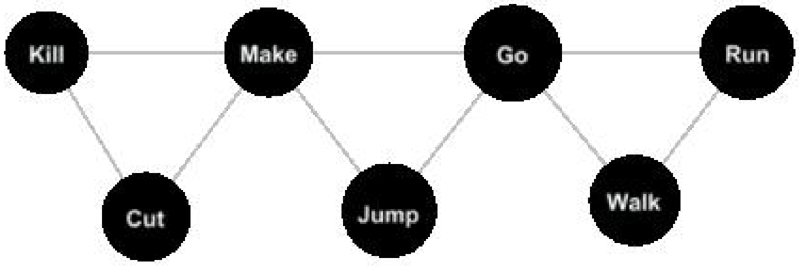
Figure 0.1: The network creation process for "Movie-1"



Fonte: Santos (2011).

The process is repeated for each movie, only adding new distinct verbs as nodes. The result for the individual data in Table 0.3 is shown in figure 3.2. Here "null" corresponds to an empty or invalid answer. One can see that the "*null*" answer was not used in the construction of the network. For each movie, the result is a clique formed by all verbs answers given for the movie. The different cliques are connected due to the pseudo-polysemy of some of the verbs, which were produced for more than one movie.

Figure 0.2: The network created from the data in Table 0.3.



Fonte: Santos (2011).

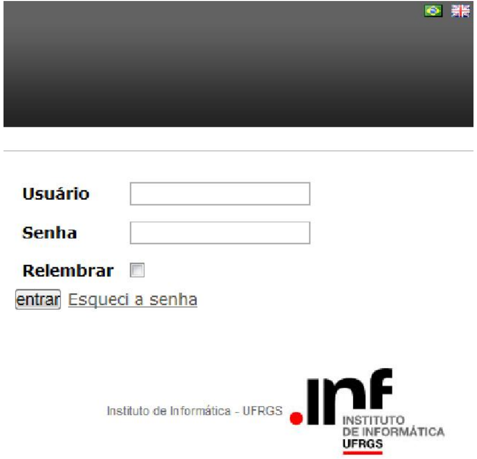
This network approach to using the linguistic production of a group to constructing the network is similar to that employed by Parente et al. (2011). We also employ the same psycholinguistic task.

1.7 Online Psycholinguistic Data Collection

An online psycholinguistic data collection software was developed in this work in order to facilitate the new psycholinguistic data to be collected for the action naming task presented in section 1.5.2. The software was developed in PHP language.

The software provides an administrative management area where new collections can be created and viewed, as well as exported to be used with the semantic network analysis software (described in section 1.8). The login and management pages can be seen in figures 3.3 and 3.4.

Figure 0.3: The login page.



Usuário

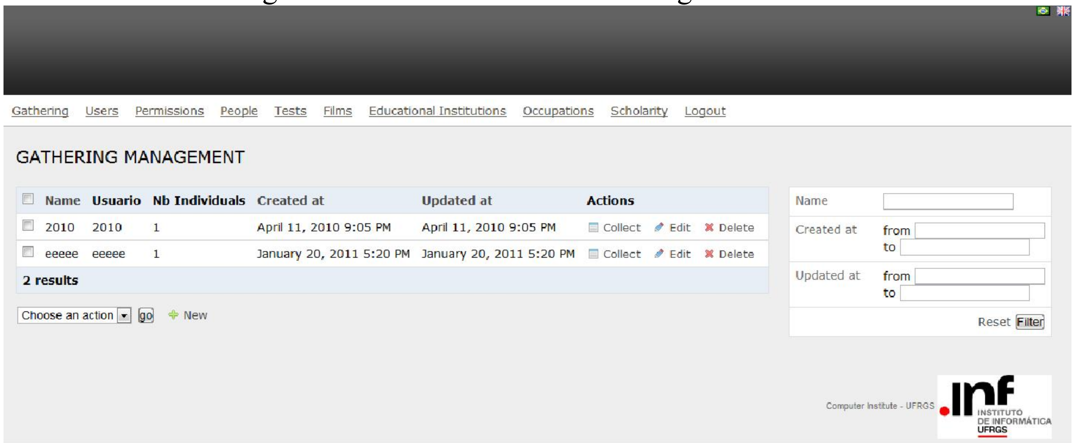
Senha

Relembrar □

entrar Esqueci a senha

Fonte: Santos (2011).

Figure 0.4: The administrative management area.



GATHERING MANAGEMENT

□	Name	Usuario	Nb Individuals	Created at	Updated at	Actions
□	2010	2010	1	April 11, 2010 9:05 PM	April 11, 2010 9:05 PM	Collect Edit Delete
□	eeeee	eeeee	1	January 20, 2011 5:20 PM	January 20, 2011 5:20 PM	Collect Edit Delete

2 results

Choose an action go New

Name

Created at from to

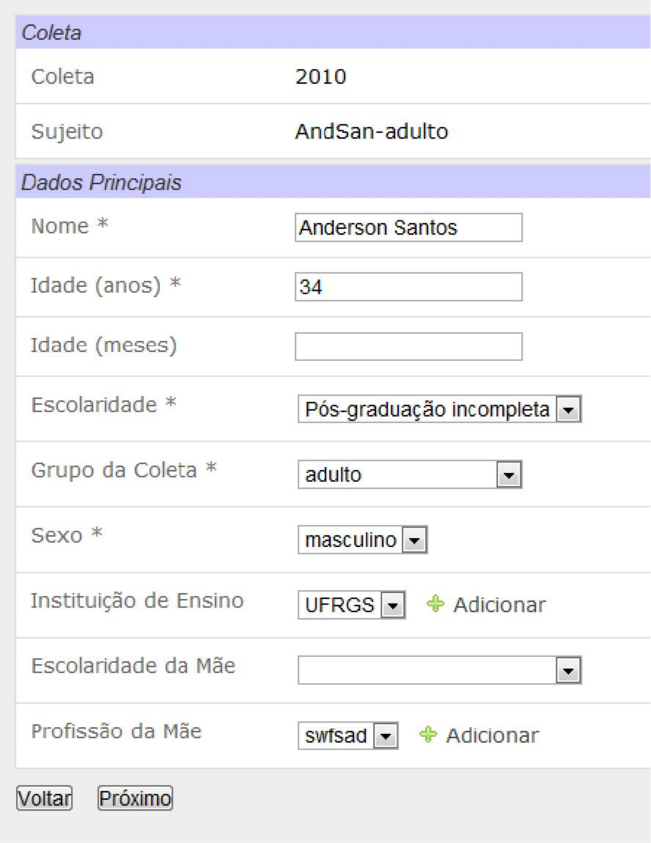
Updated at from to

Reset Filter

Fonte: Santos (2011).

The software also has a wizard for speakers' use, in order to guide the collection process. The wizard only lets users access their own profile information and task data. The screenshots of the wizard steps can be seen in figures 3.5, 3.6 and 3.7.

Figure 0.5: Collection wizard first step – Acquiring user data.



Coleta

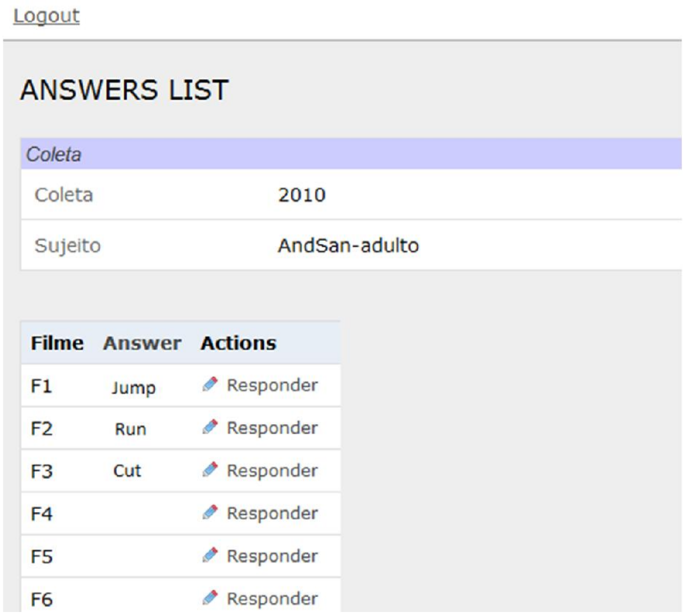
Coleta	2010
Sujeito	AndSan-adulto

Dados Principais

Nome *	Anderson Santos
Idade (anos) *	34
Idade (meses)	
Escolaridade *	Pós-graduação incompleta
Grupo da Coleta *	adulto
Sexo *	masculino
Instituição de Ensino	UFRGS
Escolaridade da Mãe	
Profissão da Mãe	swfsad

Fonte: Santos (2011).

Figure 0.6: Collection wizard second step – Answer list.



ANSWERS LIST

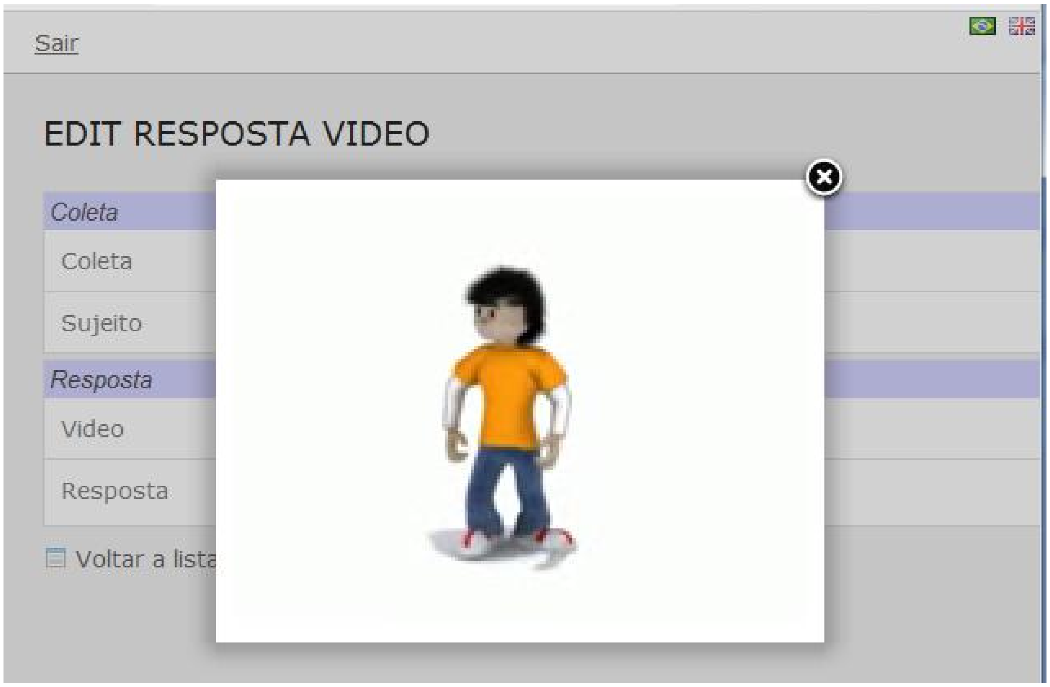
Coleta

Coleta	2010
Sujeito	AndSan-adulto

Filme	Answer	Actions
F1	Jump	Responder
F2	Run	Responder
F3	Cut	Responder
F4		Responder
F5		Responder
F6		Responder

Fonte: Santos (2011).

Figure 0.7: User viewing the movie before answering it.



Fonte: Santos (2011).

1.8 The Semantic Network Analysis

To explore and interpret networks, some complex network software tools were developed in recent years (BATAGELJ; MRVAR, 1998; SHANNON et al., 2003; ADAR, 2006; CSARDI; NEPUSZ, 2006; BASTIAN et al., 2009). An analysis of the available network analysis tools was performed in order to choose the ones that could meet our requirements:

- The software must be easy to extend (by creating plug-ins, scripts, or any sort of extensions).
- The software must be platform independent (or at least run in Windows, Mac and Linux systems).
- It must be compatible with those commonly used by researchers, as we want our contribution to be helpful to others.

As a result of this research, based on the requirements above, two toolkits were chosen: the *iGraph* library for R[9] (CSARDI; NEPUSZ, 2006; R DEVELOPMENT CORE TEAM, 2010); and the Gephi toolkit library for Java[10] (BASTIAN *et al.*, 2009). However, they have different statistical features, which make both of them very useful. In Table 0.4 there is a sample of the analysis of feature requirements of both toolkits. In this table the red symbol corresponds to the existence of the feature in the software, the green one to the absence of the feature and the coffee cup to when the feature can be easily developed in Java environment.

Table 0.4: Statistic features comparison between Gephi and iGraph libraries.

Metric	**Gephi**	**R(igraph)**
Clustering	✓	⛔
Average Degree	☕	✓
Standard Deviation Degree	☕	✓
Shortest Path (by pair of nodes)	⛔	✓
Shortest Path Average	✓	✓
Diameter	✓	✓
Closeness Centrality	✓	✓
Closeness Centrality Average	☕	✓
Closeness Centrality Standard Deviation	☕	✓
Betweenness Centrality per node	✓	✓
Betweenness Centrality Average	☕	✓
Betweenness Centrality Standard Deviation	☕	✓
Eccentricity	☕	⛔
Eccentricity Average	☕	⛔

[9] The R Project for Statistical Computing available from http://www.r-project.org/

[10] Available from http://gephi.org/

Eccentricity Standard Deviation	☕	⛔
Power law alpha fit	✓	✓
PageRank	✓	✓

Fonte: Santos (2011).

Due to the particular needs of this work of analyzing linguistic and structural aspects of the psycholinguistic data, a more specific environment was needed to allow for instance, the import of lexical resources like dictionaries and thesaurus for evaluation, input of psycholinguistic data for building networks, and so on. As the semantic network field has significant overlap with computer science, linguistics and psychology fields, a user-friendly and visually attractive interface can improve the usability of analysis and exploratory tools. As environments like R and Gephi toolkit provide well tested and technically accurate tools, they can be used as the combined basis for a user-friendly semantic network toolkit environment developed in this work.

1.8.1 The Software Architecture

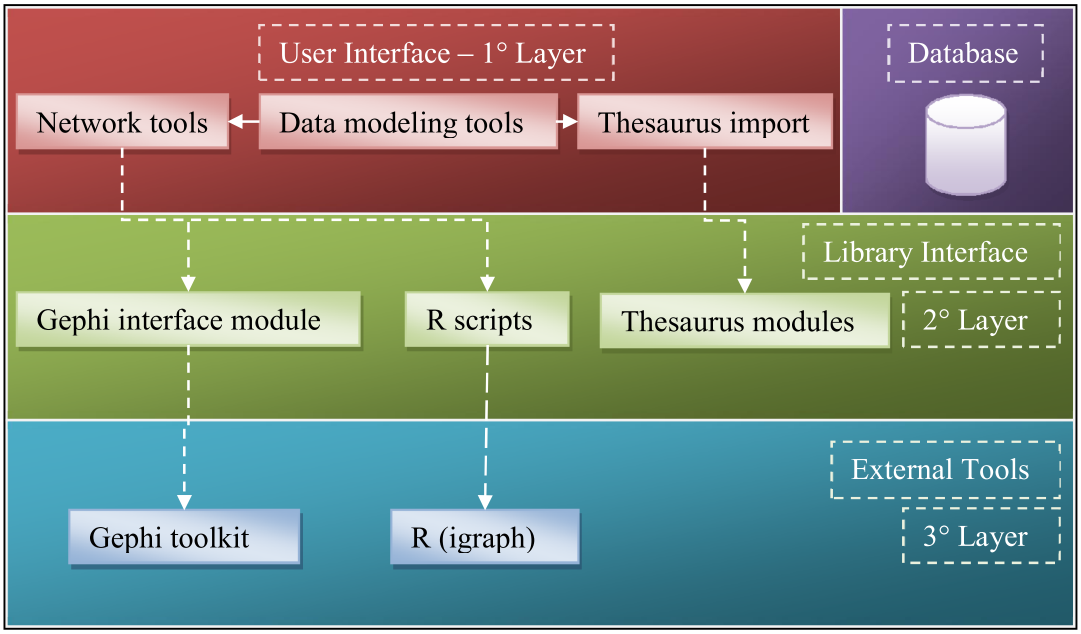
The semantic network toolkit was developed in an open source environment implemented in Pascal, using the Delphi 2010 IDE (Integrated Development Environment). The software used the Gephi and iGraph libraries in background. It has been primarily developed for Microsoft Windows operating system, but we intend to port the software to Delphi Prism IDE, so it could be compiled for .NET environment and would be able to run on Android, BSD, iOS, Linux, Mac OS X, Windows, Solaris, and Unix operating systems.

The software architecture has three main layers, as shown in figure 3.8. The first layer in figure 3.8 is the user interface also responsible for the database management. Embedded in this software layer is a database server[11]. The data psycholinguistic data collection importer's are also in the first layer and they are embedded in the Data Modeling Tools module. In order to access the Gephi toolkit library, a java application was created as one of the modules of the middle layer. Besides the R scripts (that are generated on the fly by the user interface), all other modules in this layer are self-

[11] Firebird relational database server (more info in: http://www.firebirdsql.org/).

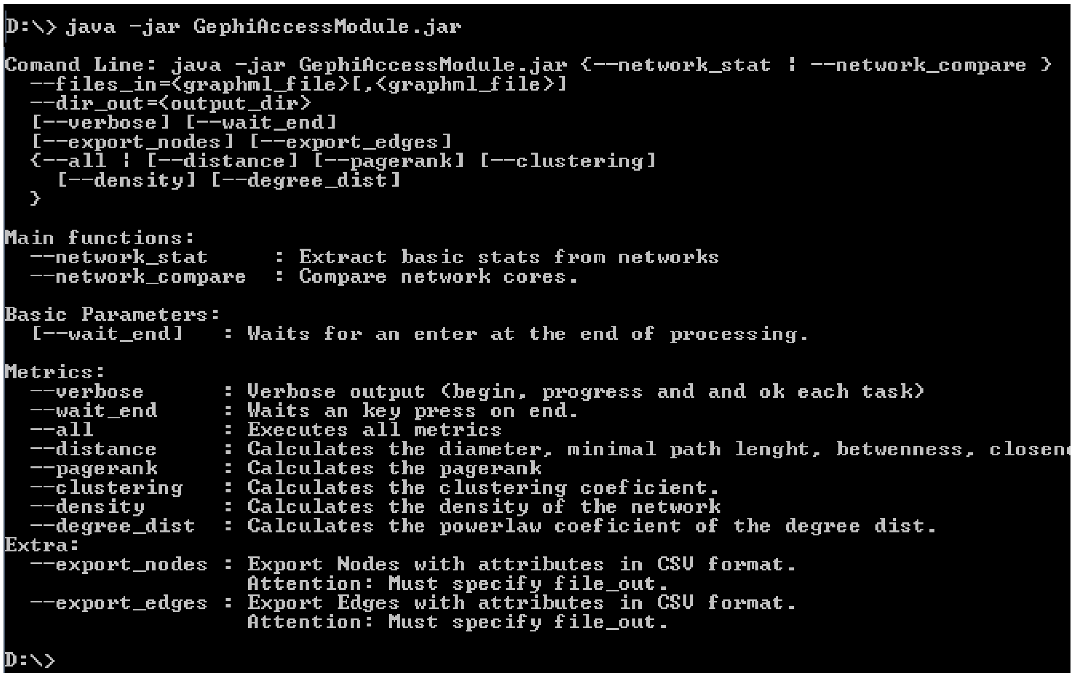
contained and can be run separately (e.g., Gephi access module output in figure 3.9). These modules are used by command-line in background when the application is running. The external softwares accessed in the background by the system are in the third layer.

Figure 0.8: The Semantic Network Toolkit Software Architecture



Fonte: Santos (2011).

Figure 0.9: Gephi access module output when used by command-line outside the semantic network toolkit.



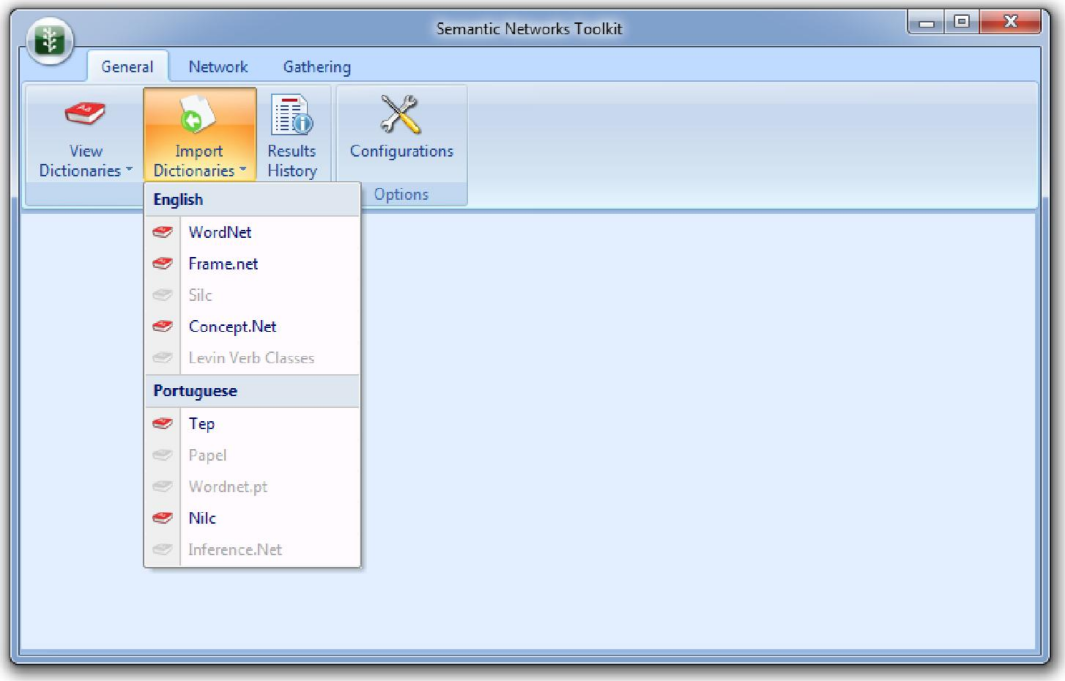
```
D:\> java -jar GephiAccessModule.jar

Comand Line: java -jar GephiAccessModule.jar {--network_stat ¦ --network_compare }
  --files_in=<graphml_file>[,<graphml_file>]
  --dir_out=<output_dir>
  [--verbose] [--wait_end]
  [--export_nodes] [--export_edges]
  {--all ¦ [--distance] [--pagerank] [--clustering]
    [--density] [--degree_dist]
  }

Main functions:
  --network_stat     : Extract basic stats from networks
  --network_compare  : Compare network cores.

Basic Parameters:
  [--wait_end]   : Waits for an enter at the end of processing.

Metrics:
  --verbose      : Verbose output (begin, progress and and ok each task)
  --wait_end     : Waits an key press on end.
  --all          : Executes all metrics
  --distance     : Calculates the diameter, minimal path lenght, betwenness, closene
  --pagerank     : Calculates the pagerank
  --clustering   : Calculates the clustering coeficient.
  --density      : Calculates the density of the network
  --degree_dist  : Calculates the powerlaw coeficient of the degree dist.
Extra:
  --export_nodes : Export Nodes with attributes in CSV format.
                   Attention: Must specify file_out.
  --export_edges : Export Edges with attributes in CSV format.
                   Attention: Must specify file_out.

D:\>
```

Fonte: Santos (2011).

1.8.2 The Software User Interface

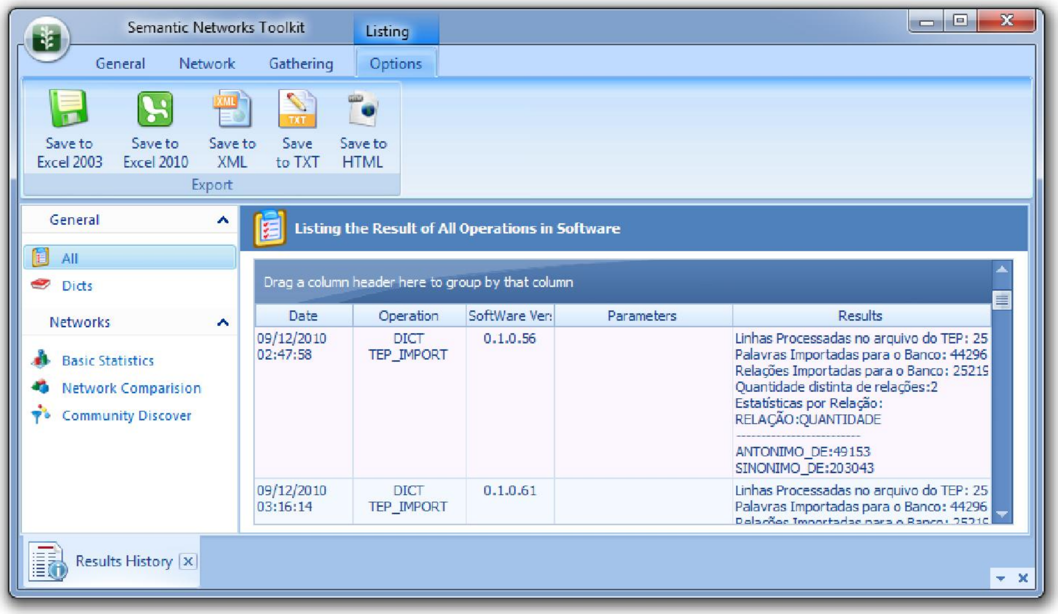
The features of the software were basically structured in three main categories: general; network and data collection. In the general category, some basic functions and reports can be accessed, as shown in figure 3.10 and 3.11. In the network category the toolkit has features that are only related to network processing (Figure 3.12). The data collection category (named gathering in the software) holds the functions related to data that are not on network format (Figure 3.13).

Figure 0.10: Semantic Network Toolkit Software – Basic Functions



Fonte: Santos (2011).

Figure 0.11: Semantic Network Toolkit Software – Reports



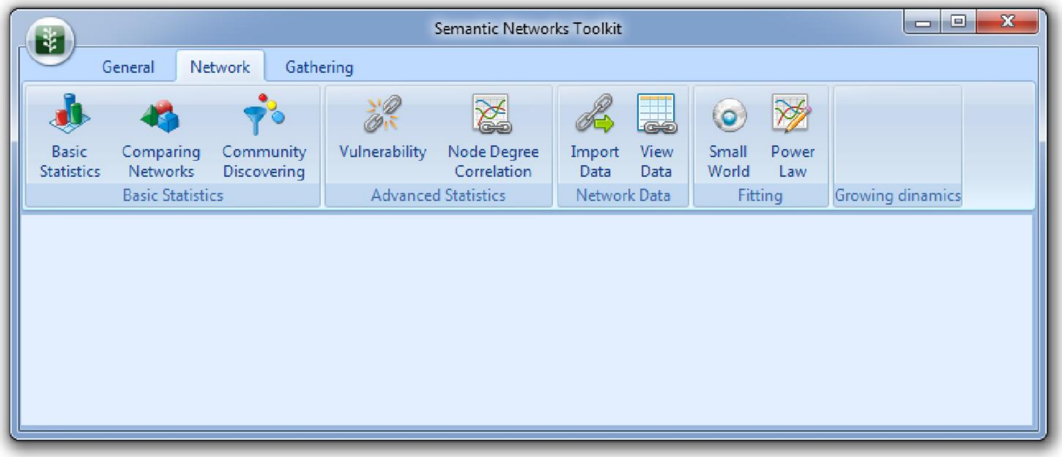
Listing the Result of All Operations in Software

Drag a column header here to group by that column

Date	Operation	SoftWare Ver:	Parameters	Results
09/12/2010 02:47:58	DICT TEP_IMPORT	0.1.0.56		Linhas Processadas no arquivo do TEP: 25 Palavras Importadas para o Banco: 44296 Relações Importadas para o Banco: 25219 Quantidade distinta de relações:2 Estatísticas por Relação: RELAÇÃO:QUANTIDADE ------------------------- ANTONIMO_DE:49153 SINONIMO_DE:203043
09/12/2010 03:16:14	DICT TEP_IMPORT	0.1.0.61		Linhas Processadas no arquivo do TEP: 25 Palavras Importadas para o Banco: 44296 Relações Importadas para o Banco: 25219

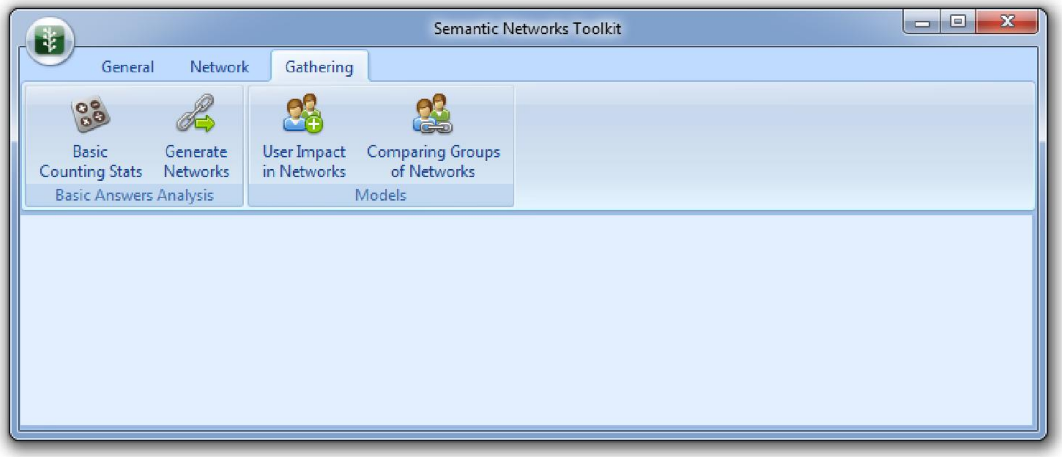
Fonte: Santos (2011).

Figure 0.12 Semantic Network Toolkit Software – Network functions



Fonte: Santos (2011).

Figure 0.13: Semantic Network Toolkit Software – Collection Models and Tools developed.



Fonte: Santos (2011).

EXPERIMENTS

The following experiments were done by testing psycholinguistic hypotheses and investigating the contributions of various factors that may have influence on the ability of language, such as frequency of words, semantic complexity, concreteness, conventionality, specificity analysis, among others.

In the first experiment, the data collected were represented as graphs, building collective semantic networks of the native speaker groups. Features were extracted from the graphs for each of the groups. Quantitative and qualitative comparisons were done in order to investigate a classifying model based on semantic network features.

In the following experiments, our focus was the construction of classifiers that capture the differences between the various profiles of speakers, and that can be used to classify new speakers according to the closest profile. We made this effort in order to help the diagnosis of diseases that affect language, such as the Alzheimer's disease.

All experiments in this chapter were implemented with the semantic network toolkit software.

1.9 Topological Analysis

For each of the elderly groups presented in section 1.5.1 (AD and HE) one semantic network was created following the methodology explained earlier, where every distinct verb uttered by a participant of the group was represented by a node in the network. A link between two nodes (verbs) was added to the network if the two verbs were uttered for the same action. The result was a clique formed by all verbs given for a movie, and the different cliques were connected due to the pseudo-polysemy of some of the verbs, which were produced for more than one movie.

A comparison of the two groups was done in terms of their structure, through topological analysis, and also of their content[12]. Table 0.1 shows some relevant topological measures (for details, see Table 0.2).

[12] This experiment was published in the interdisciplinary Workshop on Verbs conference (SANTOS *et al.*, 2010).

The results are discussed in terms of two comparisons. In the first one, we compare the semantic networks of the two groups with each other. The results are further evaluated by first determining the expected differences that would arise from a variation in the participants (using the HYA group), and then comparing them with the observed differences between the two elderly groups (AD and HE).

1.9.1 Elderly Groups

Sharing the same global features as the other language networks, the networks created for each of the elderly groups presented in section 1.5.1 show a small-world structure: they have a low average minimal path length and high-clustering coefficients when compared with random graphs of the same size and density. We used the same methodology of Steyvers & Tenenbaum (2005) in order to indentify the small-world structure present in the elderly networks.

Apart from their diameters, the two networks differ considerably in all other measures. First of all, the AD group produced more distinct verbs for describing the actions, which is reflected in a slightly larger number of nodes than the HE group and express a lower agreement for describing the actions. As a consequence, although a larger number of edges would be expected with more nodes in the AD group and their average degree of connectivity (<*k*>) in the HE one, the observed increase was considerably larger than expected.

Secondly, the average degree of connectivity, <*k*>, and standard deviation of the AD group are consistently higher than those in the HE group. One possibility for a larger *k* is the use of more polysemic verbs by the AD group – for every action in which a verb is used to describe it, it becomes connected to all other verbs also used to describe the action, forming a clique. Therefore, for each new context in which a verb is used, its degree increases by the size of the clique. If we assume that more connected verbs are also more generic, this would be consistent with the tendency of aphasic patients to use more general verbs (BREEDIN et al., 1998; THOMPSON, 2003; KIM; THOMPSON, 2004; BARDE et al., 2006; THOMPSON; SHAPIRO, 2007).

Thirdly, with a larger number of edges between the nodes and a higher average connectivity, the average minimal path length (*L*) would be expected to be smaller in the AD group than in the HE one. However, the opposite is found, which can be an indication that the differences between the two networks go beyond the use of a larger vocabulary and less agreement in the AD group, but that they are structurally different too. All this results can be seen in Table 0.1.

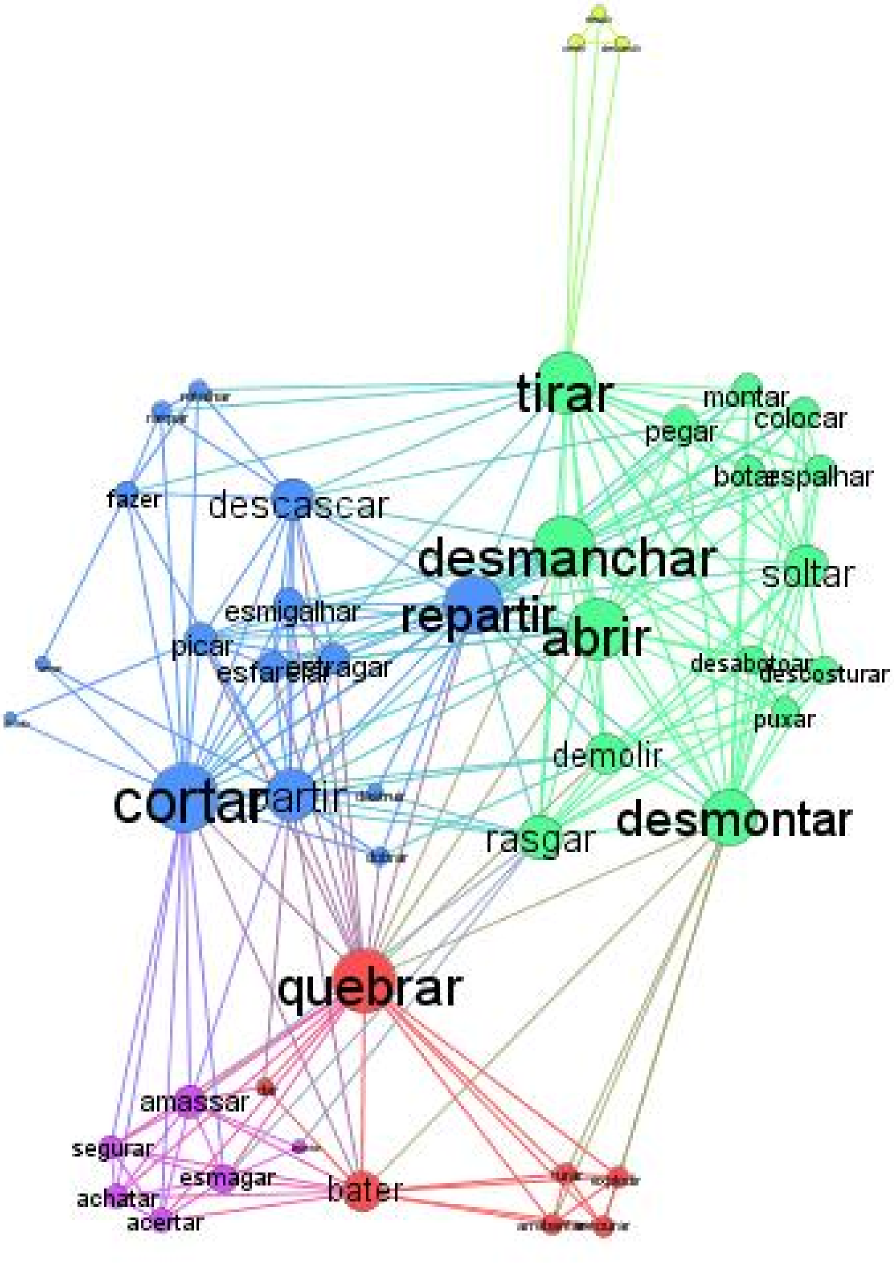
Table 0.1: A summary of AD and HE semantic network features. L_{Random} and C_{Random} are the L and C values of random graphs with the same size and density.

Variable	**AD**	**HE**
n (verbs)	46	40
m (edges)	243	140
<*k*>	10.57 (SD 6,55)	7.00 (SD 4,56)
L	1.94	1.57
D	4	4
C	.829	.789
L_{Random}	1.822	2.07
C_{Random}	.223	.175

Fonte: Santos (2011).

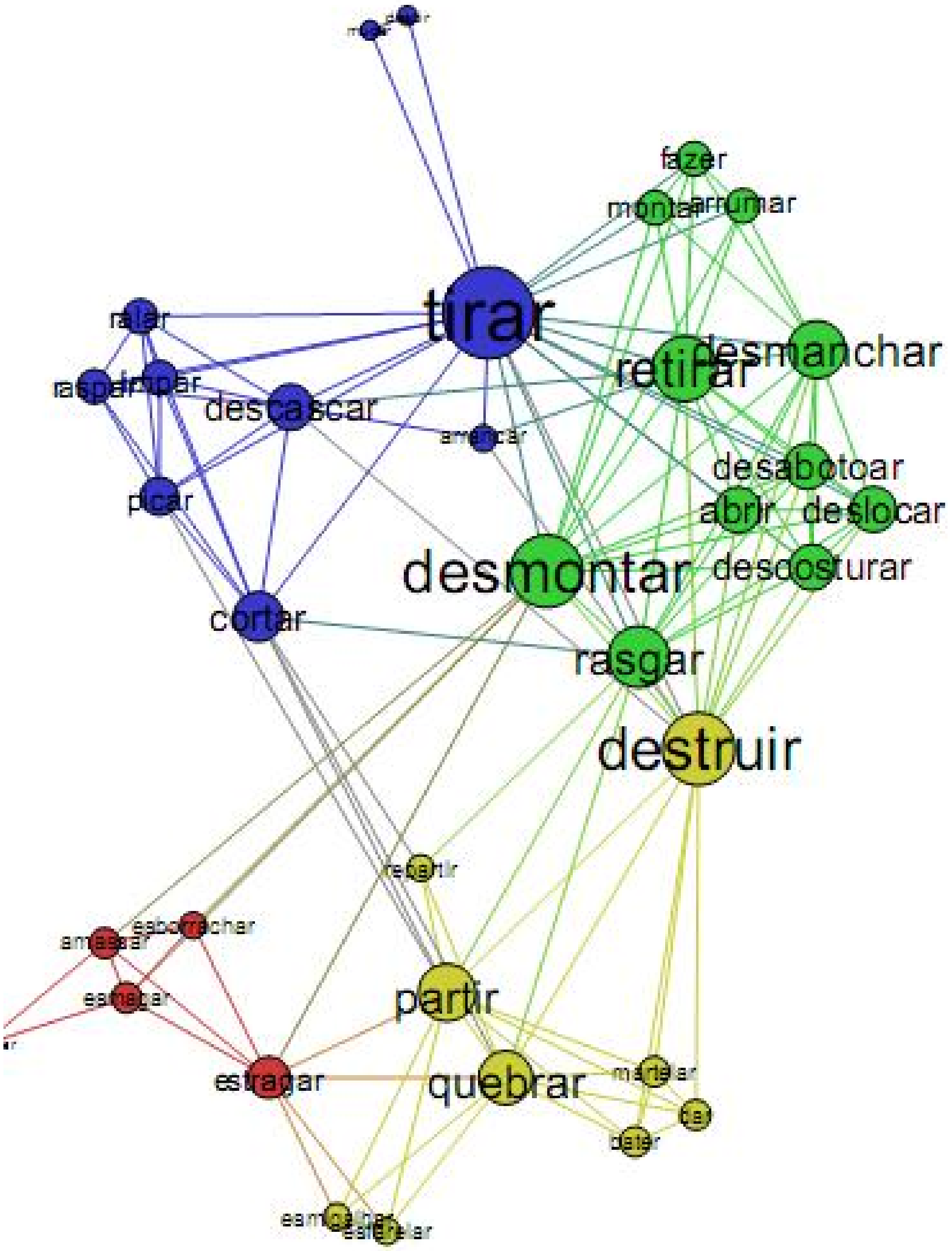
In figures 4.1 and 4.2, we can see the two networks in which the size of a node is shown in direct proportion to its degree (normalized). For visual improved comparison purpose, we used the same layout algorithm to organize the nodes of both networks - Force Atlas, which makes graphs more compact and readable (BASTIAN *et al.*, 2009). The node colors were chosen based on a community discovering algorithm (BLONDEL *et al.*, 2008). The images suggest a larger number of highly connected nodes, or hubs, in the AD network.

Figure 0.1: Network created from AD group.



Fonte: Santos (2011).

Figure 0.2: Network created from HE group



Fonte: Santos (2011).

1.9.2 Adults and Elderly Groups

In order to verify the degree of variation expected from different groups of participants, and whether this variation could explain the differences found between the two elderly groups. We also created 30 subgroups (we tested with higher number of groups and the results were similar) of 23 participants randomly selected from the 75 in the young adult group (HYA - group presented in section 1.5.1). For each subgroup, we generated a semantic network using the same method we did for the elderly groups.

Table 0.2 shows the mean and standard deviation of the topological features of the 30 groups. In addition, this table also shows the module of the difference of statistics between the AD and the HE networks. All the differences are larger than the standard deviation of the adults' samples. This indicates that intra-group variations are not enough to explain the differences found between the elderly groups. We noticed also an indication that the diameter of the network (*D*) does not seems to be a good metric to describe the differences between the semantic networks.

Table 0.2: Characterization of Adults' Sample, including the difference between Alzheimer's and control networks.

	HYA Adults' Sample		|AD-HE|
Variable	Mean	SD	
n (verbs)	38.57	1.305	6
m (Edges)	334	20.85	103
<*k*>	9.405	0.355	3.57
L	2.137	0.05	0.37
D	4.567	0.504	0
C	0.817	0.012	0.04

Fonte: Santos (2011).

With this experiment, it was possible to investigate if measures of semantic networks could reflect some changes in the semantic memory of AD patients. Some of these measures, such as the average node degree (<*k*>) are consistent with the findings of Chan, et. al. (2007), in which the authors found an increase of the number of links in the semantic networks of AD patients as the disease evolves.

By acquiring more knowledge, the semantic network of the individual will consist of less unnecessary links (SCHVANEVELDT *et al.*, 1985). If we imagine that the process of looking for a concept in the human mind could be related to a Markov decision process, and each concept (or group of concepts) represents a state, and for each link a probability to follow it. If the semantic network has more unnecessary links, greater is the probability of choosing a long way toward the concept that is being looked for. Therefore, there could be an increase in the response time of the process when looking for a concept, and there be a probability of not finding it.

1.10 A Group Classification: Linguistic and Distributional Features

To further analyze the linguistic characteristics of the semantic networks of each group, the verbs were annotated with syntactic and semantic information. The annotated data was used as the basis for a classifier that can be used to diagnose AD using the action naming task data. Using the data, a decision tree was induced by supervised classification methods[13]. We used Weka[14] (HALL *et al.*, 2009) for the construction of the classifiers, using its implementation of Bayes (Bayesian algorithms), function based algorithms such as logistic regression and Support Vector Machines (SVM), classification/regression tree algorithms and rule based algorithms. All results' accuracy statistics were generated using a ten-fold cross validation estimation.

In *k*-fold cross-validation, the original sample is randomly partitioned into *k* subsamples (KOHAVI, 1995). A single subsample data is retained for testing the model, and the remaining $k-1$ subsamples are used as training data (KOHAVI, 1995). The cross-validation process is then repeated *k* times (the *k*-folds), with each of the *k* subsamples used as the validation data (KOHAVI, 1995). The *k* results from folds are then combined to produce a single estimation (KOHAVI, 1995). The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once (KOHAVI, 1995).

Using the annotated data from the elderly (HE) and Alzheimer's (AD) groups, the classifiers were built using the features listed in Table 0.3. There were 46 instances in the data, 23 AD patients and 23 HE. For each instance we have all verb answers and for each verb answers we have more 4 attributes describing the verb (the verb in infinitive form; the conventionality degree; the specificity degree and the yahoo frequency), considering that are 17 answers we have 68 (17x4) attributes for each instance in the data. All

[13] For more about supervised classification algorithms, see (QUINLAN, 1993; ZAIANE, 2000; WATKINS; BOGGESS, 2002).

[14] Available from http://www.cs.waikato.ac.nz/ml/weka/

classifying algorithms present in Weka were used, and the validation was made using the ten-fold cross validation methodology. The best classifying algorithms were: LMT (Logistic Model Trees) (LANDWEHR *et al.*, 2005) and Naïve Bayes (JOHN; LANGLEY, 1995), which achieved 67.39% of accuracy in the induced model.

In order to analyze the contribution of each feature in this experiment, we removed one feature of the data each at a time and tested them with the Naïve Bayes algorithm; the results are in Fonte: Santos (2011).
Table 0.4. The table shows the precision reached by the classifier for each of the instances groups (AD and HE) and the total precision. The results that are worse than the baseline (the data with all features) are in red, the situations in which the algorithm reached better results are in bold and in black. As can be seen in Fonte: Santos (2011).
Table 0.4, the verb (word) does not make any difference in the results. However, the conventionality degree and the yahoo® frequency value seems to be important features since the precision decays when removed. On the other hand, with the specificity degree, the classification task seems to become harder for the algorithm, and leads to worse results.

Table 0.3: Features present for each answer in the individual's data

Features
Verb (word)
Conventionality Degree (Average)
Specificity Degree (Average)
Yahoo® Frequency

Fonte: Santos (2011).

Table 0.4: Contribution of the features for the classifier

#	Feature	Precision			Contribution
		AD	HE	Total	
0	-	60.87%	73.91%	67.39%	None
1	Verb (word)	60.87%	73.91%	67.39%	None
2	Conventionality Degree	56.52%	73.91%	65.22%	+
3	Specificity Degree	60.87%	**78.26%**	**69.57%**	-
4	Yahoo® Frequency	60.87%	69.57%	65.21%	+

Fonte: Santos (2011).

In the next experiment, we investigate the contribution of the topological features for the classification task.

1.11 A Group Classification: Network Impact Analysis

In order to determine the homogeneity of the groups, we examine how the addition of the linguistic production of a new individual impacts on the structural features of the networks, assuming that the greater the impact on the group, the smaller the chance that the individual belongs to that group. If the results confirm this hypothesis, we want to further investigate the use of these features for building classifiers.

The impact of an individual is calculated as follows: for each individual, we are going to generate two networks: one with and one without him. The differences between the topological measures of these two networks are used to annotate the linguistic production of the participant. The impact is calculated for both AD and HE networks, so that for each participant, 4 networks are generated: AD-complete, AD-without, HE-complete, HE-without. Moreover, for maintaining group sizes constant, the impact of a participant in an external group is measured after excluding a participant from this group. In addition, to ensure robustness during exclusion, as there are 23 participants per group, we use the average measures obtained with the 23 groups formed by excluding a different participant of the group.

The annotated linguistic profiles are subsequently used for building classifiers. Therefore, we are going to attain a profile for each individual, as shown in the example in Table 0.5. As can be seen in the table, the individual profile contains his impact on his group and on all others.

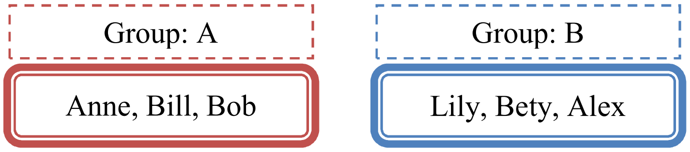
Table 0.5: The individual profile with his network impact features

Individual Profile		
(AD-complete) – (AD-without)	Nodes	+1
	Edges	-5
	C	-0.234
	L	+0.57
	…	…
(HE-complete) – (HE-without)	Nodes	+3
	Edges	+15
	C	+0.454
	L	-0.643
	…	…

Fonte: Santos (2011).

For didactic purposes, we are going to present a toy example, using two small groups of hypothetical individuals for the step-by-step calculation of the impact measure. Groups A and B and their individuals are illustrated in figure 4.3. Each individual is modeled in practice as a vector containing his/her linguistic production for the action naming task data, as shown in the example in Table 0.3, but for reasons of clarity, in this example, we represent the vector using the participant's names.

Figure 0.3: Toy example of groups' semantic networks.



Fonte: Santos (2011).

First, the algorithm generates one network for each group, extracting the topological features (explained in section 1.9.1). We are going to name these networks group A and group B.

Now, for each individual, there are two phases. In the first phase, the impact of the individual on his original group is calculated, and in the second phase, his impact on the external groups is computed. The steps of the first phase are:

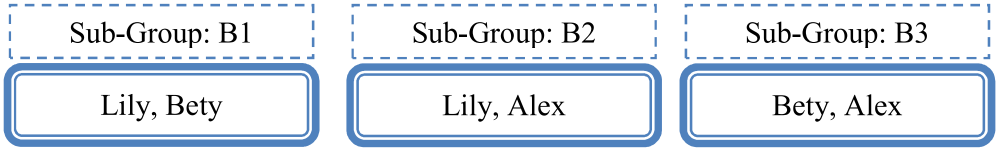
- Remove the individual from his original group.
- Generate the new network without the individual and extract the network features.
- Calculate the difference between the network with the individual and the new network without him.

To generate the first phase data for individual Anne, who belongs to group A, a network formed with only Bill and Bob's data is generated and its features are extracted. The difference between network A and the network formed only by Bill and Bob's data will be computed and saved in Anne's profile.

In the second phase Anne is added to external group B which already contains 3 participants. However, if we just insert Anne in group B, we are going to compare features of different group sizes. In order to avoid this unbalanced comparison, *n* subgroups of B will be generated; each with the exclusion of a different participant where *n* is the number of elements in group B (3 in this case). The resulting sub-groups of the example are shown in figure 4.4, each with *n-1* elements. Now, the algorithm can add individual Anne of group A to each of the B sub-groups.

For each external B sub-group, a network with and without Anne is generated. The average differences in graph features between the networks with and without Anne are computed and saved in Anne's profile.

Figure 0.4: The sub-groups of group B



Fonte: Santos (2011).

1.11.1 Results

The algorithm of network impact analysis was executed to generate the profiles of the groups (AD and HE). The network features used to create the model were the same features of the experiment presented in section 1.9 (repeated below for convenience) with a few more:

Feature
n (verbs)
m (Edges)
<*k*>
L
D
C
Density (relation between number of nodes and edges)
Average Betwenness (see Section 1.1.1)
Average Closeness (see Section 1.1.1)
Average Eccentricity (see Section 1.1.1)
Average PageRank (see Section 1.1.1)

The data for each participant is represented by a vector, or profile, containing each of these measures and information about group membership, for building the classifiers with the Weka mining tool. A sample of the profile vector is shown in Table 0.6. All classifying algorithms were tested using ten-fold cross-validation test methodology. The best results were achieved by two algorithms, one decision-tree classifier and one classifier based on rule induction.

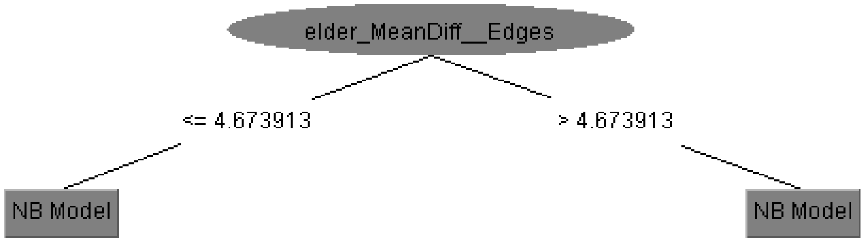
Table 0.6: A sample of the profile vector used to create the classifier

OwnGroup	alzheimer_MeanDiff__Nodes	alzheimer_MeanDiff__Edges	alzheimer_MeanDiff_Avg_Path_Lenght	…
alzheimer	3	38	-0,06905	…
alzheimer	0	10	-0,02319	…
alzheimer	1	5	0,003645	…
alzheimer	1	15	-0,00747	…
alzheimer	1	15	-0,00646	…
alzheimer	0	2	-0,00676	…

Fonte: Santos (2011).

The best decision-tree algorithm was the NBTree algorithm (KOHAVI, 1996). The algorithm achieved 82.61% of accuracy in the generated model with the decision tree shown in figure 4.5. The NBTree algorithm (KOHAVI, 1996) is a mixed algorithm which induces a hybrid of decision-tree classifiers and NaiveBayes (JOHN; LANGLEY, 1995): the decision-tree nodes split as regular decision-trees, but the leaves contain Naive-Bayesian (JOHN; LANGLEY, 1995) classifiers, as can be seen in figure 4.5.

Figure 0.5: The tree generated by NB-Tree (KOHAVI, 1996) algorithm.



Fonte: Santos (2011).

The model misclassifications can be seen through the confusion matrix presented in Table 0.7. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class reflecting if the system is confusing two classes (i.e. commonly mislabeling one as another). In Table 0.7 two Alzheimer's individuals were misclassified as healthy elders, and six elders were also misclassified as Alzheimer's.

Table 0.7: Confusion matrix of NB-Tree generated model using the data from the network impact model.

Classified as =>	Alzheimer's	Elder

Alzheimer’s	21	2
Elder	6	17

Fonte: Santos (2011).

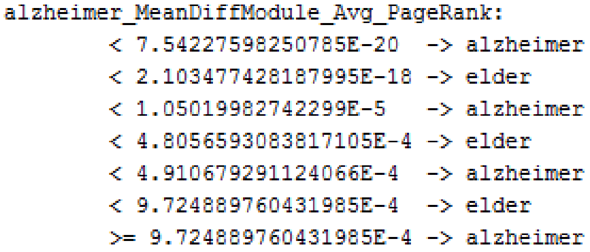
The best rule induction algorithm was the OneR (HOLTE, 1993). The OneR algorithm only uses one attribute to build the predictive model. The rules are created based on the selected attribute. The algorithm reached 86.96% of accuracy in the created model. The attribute chosen by the algorithm was the module of the average PageRank feature difference in the Alzheimer’s network. That indicates that the average impact of the PageRank on the Alzheimer’s network is the one of most distinguishing features between AD and HE individuals; the rules created by the algorithm are shown in figure 4.6. These rules divide the PageRank spectrum into ranges which are associated with each of the groups. For example, if the value of the measure is between 2.1035 and 7.5423, the classifier predicts that the participant belongs to the AD group. The confusion matrix is in Table 0.8.

Table 0.8: The confusion matrix of OneR (HOLTE, 1993) generated model using the data from the network impact model.

Classified as =>	**Alzheimer’s**	**Elder**
Alzheimer’s	22	1
Elder	5	18

Fonte: Santos (2011).

Figure 0.6: The rules generated by OneR (HOLTE, 1993) algorithm.



```
alzheimer_MeanDiffModule_Avg_PageRank:
	< 7.54227598250785E-20  -> alzheimer
	< 2.103477428187995E-18 -> elder
	< 1.05019982742299E-5   -> alzheimer
	< 4.8056593083817105E-4 -> elder
	< 4.910679291124066E-4  -> alzheimer
	< 9.724889760431985E-4  -> elder
	>= 9.724889760431985E-4 -> alzheimer
```

Fonte: Santos (2011).

The results of both algorithms provided the best speaker groups diagnosis from the options investigated in this work. The results also show that complex network metrics play an important role in this process.

As some important cues and features masked by the averaging process, in order to explore this network impact model without using the averaging process, we consider alternative configurations in the next experiment.

1.12 Network Impact Analysis Maximizing the Data

In order to maximize the data and to find if the averaging process in the external networks impact causes impact on the accuracy of the classifiers, we are going to consider alternative configurations.

For this experiment we used the same individual profile structure described in 1.11, but instead of using the average impact on the external network as a single feature, we used impact information for each individual network as features. For instance, for a classification in 2 groups, the vector now contains only 2 sets of features: the set of the impact on group 1 and the set of the impact on group 2. For the impact on the original group, we also followed the same methodology when extracting the individual impact of his original network (first phase of the algorithm in 1.11). In the second phase, the impact on each of the different networks corresponds to a new instance.

For example, given the toy examples shown in figure 4.4, the vector contained 2 sets of impact features and there were 3 instances for each participant. For instance, for Anne, each instance contains her impact on group A and on one of the B subgroups. For individual Anne, a profile was generated for each B sub-group. All profiles have Anne's impact on her original group, and Anne's impact on the respective B sub-group. The resulting data for the example is shown in Table 0.9. The last attribute (that is not shown in Table 0.9) corresponds to the class information of the individual and is used only for training.

Table 0.9: Data result from the example

Individual	A impact	B impact
Anne_vs_B1	Anne's impact on group A	Anne's impact on group B1
Anne_vs_B2	Anne's impact on group A	Anne's impact on group B2
Anne_vs_B3	Anne's impact on group A	Anne's impact on group B3
Bill_vs_B1	Bill's impact on group A	Bill's impact on group B1
Bill_vs_B2	Bill's impact on group A	Bill's impact on group B2
Bill_vs_B3	Bill's impact on group A	Bill's impact on group B3
Bob_vs_B1	Bob's impact on group A	Bob's impact on group B1
Bob_vs_B2	Bob's impact on group A	Bob's impact on group B2
Bob_vs_B3	Bob's impact on group A	Bob's impact on group B3
Lily_vs_A1	Lily's impact on group A1	Lily's impact on group B

Lily_vs_A2	Lily’s impact on group A2	Lily’s impact on group B
Lily_vs_A3	Lily’s impact on group A3	Lily’s impact on group B
Bety_vs_A1	Bety’s impact on group A1	Bety’s impact on group B
Bety_vs_A2	Bety’s impact on group A2	Bety’s impact on group B
Bety_vs_A3	Bety’s impact on group A3	Bety’s impact on group B
Alex_vs_A1	Alex’s impact on group A1	Alex’s impact on group B
Alex_vs_A2	Alex’s impact on group A2	Alex’s impact on group B
Alex_vs_A3	Alex’s impact on group A3	Alex’s impact on group B

Fonte: Santos (2011).

1.12.1 Results

We used the different configuration of the features for the HE and AD groups, as the input for building the classifiers implemented in the Weka[15] (HALL *et al.*, 2009) software. The algorithms were evaluated using ten-fold cross-validation. Almost all of them reached more than 90% of accuracy.

The best resulting model was generated by the Random Forest (STATISTICS; BREIMAN, 2001) decision-tree based algorithm: 97.35% of precision. The confusion matrix showing the errors of the classifier can be seen in Table 0.10. There are more instances of data than the results shown earlier; as each individual generates 23 instances of data.

For the error analysis, we explored the 23 misclassification instances of the Alzheimer’s group. All of them were generated by the same individual, and the reason for the misclassification was that the individual does not cause any impact on the HE sub-group networks, as it does not make any changes on the set of nodes and edges. Further comparison between his answer set and those of the HE group revealed that although his answer set is unique, the reason why it causes no impact on the HE groups is that he always gives one of the most frequent responses for each video in the HE group. Further confirmation of the AD diagnosis for that patient is necessary.

Table 0.10: The confusion matrix of Random Forest (STATISTICS; BREIMAN, 2001) generated model using the data from the improved network impact model.

Classified as =>	**Alzheimer’s**	**Elder**

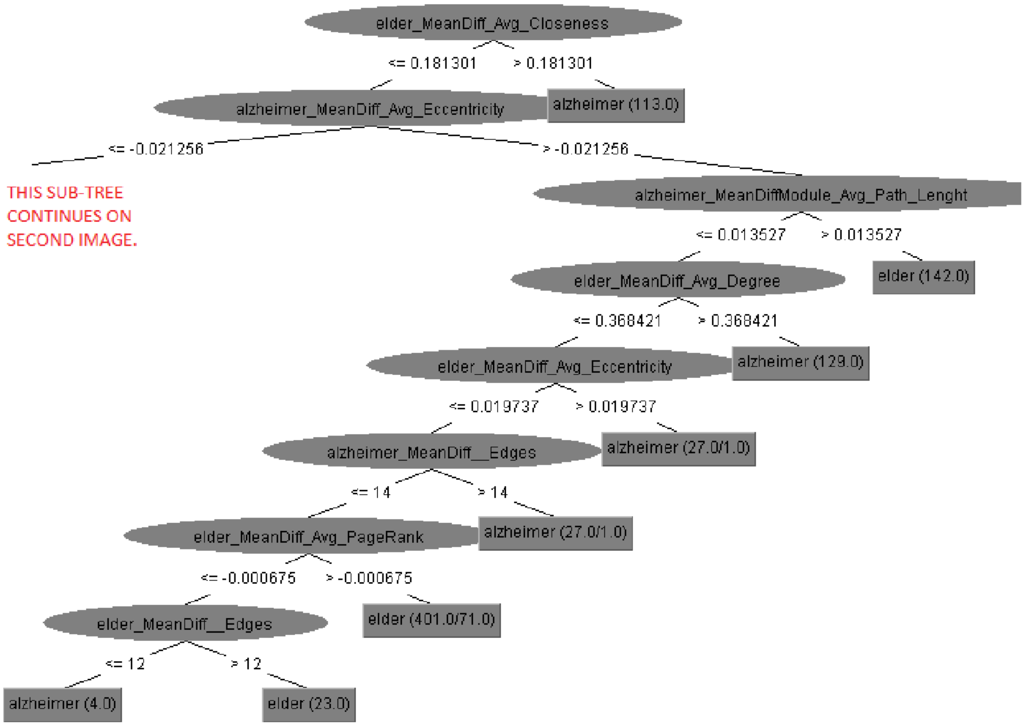
[15] Weka version: 3.6.4

Alzheimer’s	506	23
Elder	5	524

Fonte: Santos (2011).

Since the Random Forest algorithm implementation in Weka (HALL *et al.*, 2009) does not show the generated trees for a qualitative discussion of the results, we are going to use the tree from the J48 (HOLTE, 1993) algorithm which resulted in the second best precision of the decision-tree classifying algorithms: 92.16%. The tree was split into two figures due to size limitations, and is shown in figure 4.7 and 4.8.

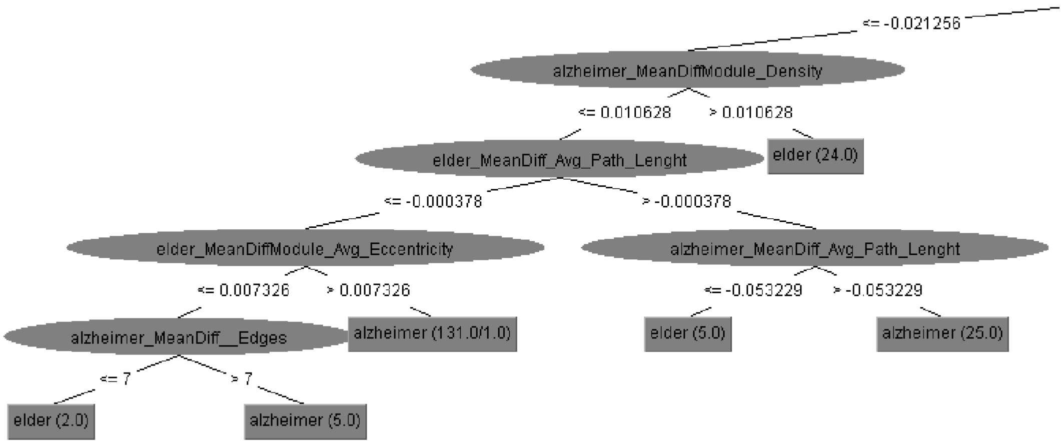
Figure 0.7: The decision tree model generated by the J48 algorithm (first part).



THIS SUB-TREE CONTINUES ON SECOND IMAGE.

Fonte: Santos (2011).

Figure 0.8: The decision tree model generated by the J48 algorithm (second part).



Fonte: Santos (2011).

As can be seen in figure 4.7 and 4.8, the tree has 14 leaves, with 27 decision-nodes and a height size of 8. Considering that it represents 1058 instances of data and was tested using a ten-fold evaluation method, it is not an over fitting decision tree. The confusion matrix of the decision tree can be seen in Table 0.11.

Table 0.11: The confusion matrix of J48 algorithm (HOLTE, 1993) generated model using the data from the improved network impact model.

Classified as =>	**Alzheimer's**	**Elder**
Alzheimer's	450	79
Elder	4	525

Fonte: Santos (2011).

As can be seen in Table 0.10 and Table 0.11, only a few HE instances were misclassified. However, considering that each individual generates 23 instances to be classified by the model, when predicting the individual group, a voting procedure based on the outcome of instances can be used to predict the individual's class. If we consider that a single senior citizen generated instances that were misclassified by both models (in the worst case), this individual would still be classified correctly by the voting process. Therefore, the model presented in this section has not produced any false positive error (e.g., any elder misclassified as Alzheimer's).

1.12.2 Refining the model data

In order to find one of the set of attributes that could lead to a better accuracy on the results, and also to determine the contribution of each of the features in the classification task, we used a genetic search algorithm proposed by Goldberg (1989) to generate the attribute sets.

A genetic algorithm (GA) is a search technique used to find approximate solutions in optimization problems and search. Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, natural selection, and recombination (or crossing over).

Genetic algorithms are implemented as a simulation in which a population of abstract representations of the solution is selected in search of better solutions. The evolution usually starts from a set of randomly created solutions, and is carried through generations. In each generation, the suitability of each solution is evaluated in the population, some individuals are selected for the next generation, and are mutated or recombined to form a new population. The new population is then used as input for the next iteration.

Each set of attributes generated by the genetic algorithm was evaluated using the J48 (HOLTE, 1993) classifier and ten-fold cross validation. The parameters used in the genetic algorithm are shown in Table 0.12. The solution is represented as a set of Boolean values; each value represents one of the attributes present in the individual's profile. So, the genetic algorithm tried to find the set of attributes that can produce the best results and, therefore finding the most representative attributes in the individual's profile.

Table 0.12: Parameters used in the genetic algorithm (GOLDBERG, 1989) to perform the best attribute set search.

Parameter	**Value**
Population Size	20
Generations	20
Cross-over Probability	60%
Mutation Probability	3,3%
Start Population	Random

Fonte: Santos (2011).

The result of the search for the best set of attributes is shown in Table 0.13. The decision tree generated by the J48 algorithm using only the attribute set shown in Table 0.13 is a little bit different and bigger than the one generated using all attributes: with 22 leaves and 43 decision-nodes. However, the accuracy achieved is very similar: 92.53%. The resulting decision tree is shown in figure 4.0 and 4.10.

.

Table 0.13 also shows the impact on the accuracy and on the size of the tree if we withdraw one attribute from the set. Most of the attributes does not make a considerable change on the accuracy. However, without the *AD_Diff_Degree_Dist* and *HE_Diff_Degree_Dist* attributes, the size of the tree had a substantial reduction with a very small impact on the accuracy.

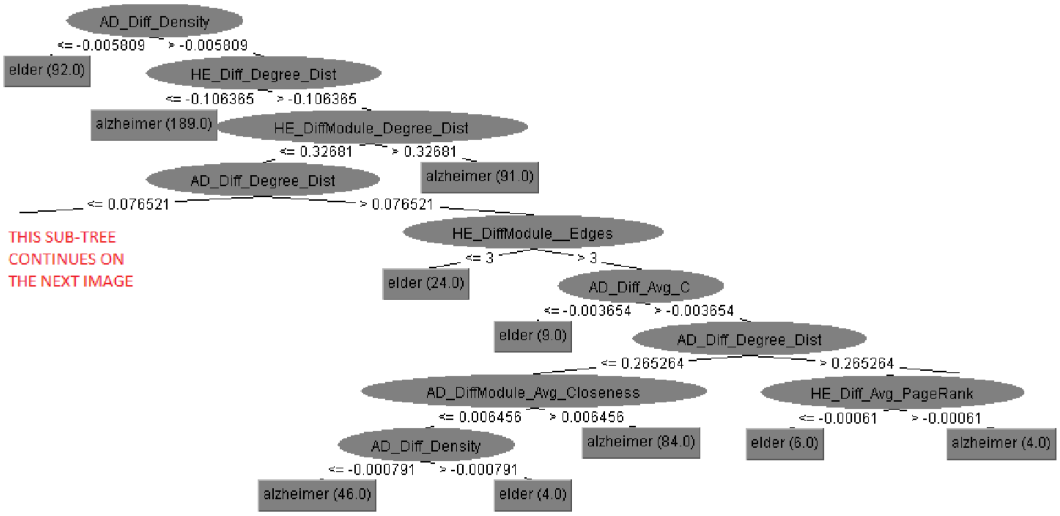
The removal of the *AD_DiffModule_Avg_Betwenness* attribute improved the accuracy and also reduced the tree size.This is probably due to the fact that, for the classifying algorithm, the attribute appears to be a good choice; however, choosing it leads to the necessity of creating a big sub-tree under it (as can be seen in figure 4.10). When the attribute is removed, the bad choice is avoided. This problem often occurs with classifying algorithms.

Table 0.13: The result of the search for the best set of attributes using the genetic algorithm for search and the attribute impact if it was ignored (GOLDBERG, 1989) and the J48 (HOLTE, 1993) classifier for evaluation.

Attribute Set	**Ignoring the attribute**		
	Accuracy	**Tree size**	**Number of leaves**
AD_Diff_Avg_C	92.34%	41	21
AD_Diff_Density	92.34%	29	15
AD_Diff_Degree_Dist	92.25%	33	17
AD_DiffModule__Nodes	92.44%	43	22
AD_DiffModule_Avg_C	92.44%	43	22
AD_DiffModule_Avg_Betwenness	92.63%	35	18
AD_DiffModule_Avg_Closeness	92.34%	35	18
HE_Diff_Avg_C	92.53%	39	20
HE_Diff_Degree_Dist	92.16%	31	16
HE_Diff_Avg_PageRank	92.53%	43	22
HE_DiffModule__Edges	92.25%	41	21
HE_DiffModule_Degree_Dist	92.25%	43	22
HE_DiffModule_Avg_Degree	92.44%	43	22
-	**92.53%**	**43**	**22**

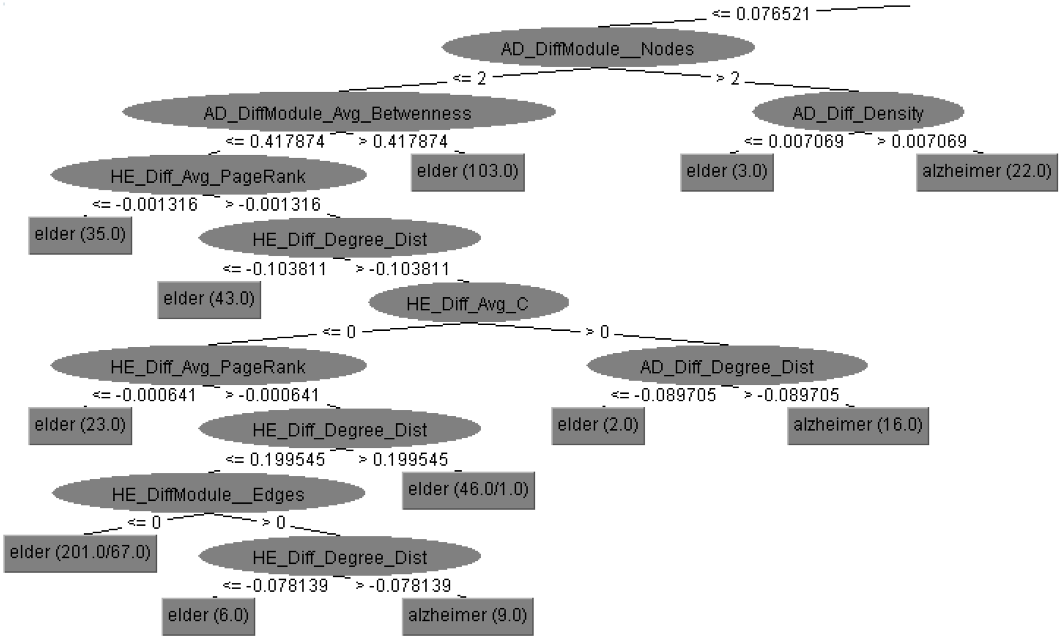
Fonte: Santos (2011).

Figure 0.9: The decision tree model generated by the J48 algorithm using the attribute set in Table 0.13 (first part).



Fonte: Santos (2011).

Figure 0.10: The decision tree model generated by the J48 algorithm using the attribute set in Table 0.13 (second part)



Fonte: Santos (2011).

Nine of the thirteen attributes (*C*, density, degree distribution, the average degree and the number of nodes and edges) selected by the algorithm were evaluated in our first experiment, and therefore the automatic selection is consistent with findings of our topological analysis in section 1.9.

CONCLUSIONS AND FUTURE WORKS

In this work we presented an investigation of the lexical organization of verbs in the context of Alzheimer's disease patients. We looked at characteristics of the semantic network of verbs produced by AD patients in an action naming task, compared with that of healthy individuals. We analyzed the collective semantic networks using statistical and topological analysis, and found interesting divergences. In particular there seemed to be less agreement among the AD patients for the lexical choice to describe a given action. In addition, there were also indications of structural differences between the networks which may arise from modifications in the lexical organization caused by AD.

This work also proposed a new approach to the classification based on the impact on structural features of collective semantic networks. The classifier was evaluated and it showed to be very accurate against our baseline comparison strategy (97% vs. 67%). It is noteworthy that the proposed classifier uses only part of the data present in the baseline strategy.

We must take into account that the results obtained in this work made very limited use of linguistic information. Therefore, even more data from the action naming task can be used, such as response time or more linguistic information about the verbs such as their syntactic and semantic aspects, in order to improve the results.

These experiments confirm that computational investigations of psycholinguistic data and concepts in terms of machine learning classifiers and complex network theories can provide a good basis for pathologies investigation. More studies need to be conducted, but this work already shows that some indications of changes in terms of lexical-semantic access can be found in the task of naming actions between the group of AD and healthy elders. It is important to consider that in addition to the linguistic response, the task also involves visual perceptual skills as the response requirements (action recognition).

Another contribution of this work is that we developed a web online psycholinguistic data collection tool for collecting and analyzing psycholinguistic data from groups of speakers. This work also provided a new open source semantic network analysis tool that is available to be used or improved by the academic community.

1.13 Future Works

For future works we envisage more data collection of different groups of speakers and using different types of verbs (not only the division and destruction verbs that were used in this work). Larger data sets will allow other types of analysis and further corroboration of the results obtained.

We also plan to analyze qualitative differences among hubs between the networks. Finally we intend to inspect other statistical features of complex networks, particularly those related to network vulnerability (CRIADO *et al.*, 2005), that are associated to network performance and helps to measure the response of complex networks subjected to attacks on vertices and edges.

We aim to investigate language acquisition, including the analysis of psycholinguistic data from the following native speaker groups:

- Children (2 to 3.3 years)
- Children (4 to 6 years)
- Young adults
- Healthy elders
- Elderly people with pathologies.

Moreover, we aim to develop tools and resources to automatically annotate existing corpora of child-produced or child-directed speech, such as Florianopolis (SCLIAR-CABRAL, 2004) for Portuguese and SACHS for English, which are part of the CHILDES database (MACWHINNEY, 2000). These corpora contain plain texts, and will be automatically annotated with distributional and syntactic information from parsers, such as PALAVRAS (BICK, 2000) for Portuguese and RASP (BRISCOE *et al.*, 2006) for English. These data will also be represented in the form of graphs for topological analysis. We aim to find more clues about the factors that could impact on language acquisition, organization and dissolution.

REFERENCES

ADAR, E. GUESS: a language and interface for graph exploration. Proceedings of the SIGCHI conference on Human Factors in computing systems. **Anais** p.791-800, 2006. New York, NY, USA: ACM. Disponível em: <http://doi.acm.org/10.1145/1124772.1124889>.

ALZHEIMER'S DISEASE INTERNATIONAL. **World Alzheimer Report 2009: Executive Summary**. 2009.

ANDERSON, J. R. **Learning and memory: An integrated approach**. 2nd ed. New York: Wiley, 2000.

BARABÁSI, A. L. Statistical mechanics of complex networks. **Reviews of Modern Physics**. v. 74, p.47-97, 2002.

BARABÁSI, A. L.; ALBERT, R. Emergence of Scaling in Random Networks. **Science**, v. 286, p. 509-512, 1999.

BARDE, L. H. F. SCHWARTZ, M. F.; BORONAT, C. B. Semantic weight and verb retrieval in aphasia. **Brain and Language**, v. 97, n. 3, p. 266-278, 2006. Disponível em: <http://www.sciencedirect.com/science/article/B6WC0-4HTBKT6-1/2/846238c63bb6f5fc039a45deaef3b778>.

BASTIAN, M. HEYMANN, S.; JACOMY, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. **International AAAI Conference on Weblogs and Social Media**, p. 361-362, 2009. AAAI. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/download/154/1009>. .

BATAGELJ, V.; MRVAR, A. PAJEK -- Program for large network analysis. ,1998. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.27.9156>.

BELL, E. E. CHENERY, H. J.; INGRAM, J. C. Semantic priming in Alzheimer's dementia: evidence for dissociation of automatic and attentional processes. **Brain and language**, v. 76 (2), p. 130-144, 2001.

BERCHTOLD, N. C.; COTMAN, C. W. Evolution in the Conceptualization of Dementia and Alzheimer's Disease: Greco-Roman Period to the 1960s. **Neurobiology of aging**, May. 1998. Elsevier. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0197458098000529?showall=true>.

BICK, E. **The Parsing System" Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. Aarhus University Press, 2000.

BLONDEL, V. D. GUILLAUME, J. L. LAMBIOTTE, R.; LEFEBVRE, E. Fast unfolding of communities in large networks. **Journal of Statistical Mechanics: Theory and Experiment**, v. 2008, p. P10008, 2008. IOP Publishing. Disponível em: <http://iopscience.iop.org/1742-5468/2008/10/P10008>. Acesso em: 2/2/2011.

BOUTTIER, J. FRANCESCO, P. D.; GUITTER, E. Geodesic distance in planar graphs. **Nuclear Physics B**, v. 663, n. 3, p. 535-567, 2003. Disponível em: <http://www.sciencedirect.com/science/article/B6TVC-48KW72R-1/2/ae6bcc93787b3ace9b7b6ff265631052>.

BREEDIN, S. D. SAFFRAN, E. M.; SCHWARTZ, M. F. Semantic factors in verb retrieval: an effect of complexity. **Brain and language**, v. 63, n. 1, p. 1-31, 1998. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/9642018>.

BRISCOE, E. CARROLL, J.; WATSON, R. The Second Release of the RASP System. **Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions**, 2006. Sydney.

CADINI, F. ZIO, E.; PETRESCU, C. A. Using centrality measures to rank the importance of the components of a complex network infrastructure. **Critical Information Infrastructure Security**, p. 155–167, 2009. Springer. Disponível em: <http://www.springerlink.com/index/f287188641922n96.pdf>. Acesso em: 11/3/2011.

CHAN, A. S. BUTTERS, N. PAULSEN, J. S. et al. An assessment of the semantic network in patients with alzheimer's disease. **J. Cognitive Neuroscience**, v. 5, n. 2, p. 254-261, 1993. Cambridge, MA, USA: MIT Press. Disponível em: <http://portal.acm.org/citation.cfm?id=1326977.1326986>. .

CHAN, A. S. BUTTERS, N.; SALMON, D. P. The deterioration of semantic networks in patients with Alzheimer's disease: a cross-sectional study. **Neuropsychologia**, v. 35, n. 3, p. 241-248, 1997. Disponível em: <http://www.biomedsearch.com/nih/deterioration-semantic-networks-in-patients/9051673.html>.

CHAN, A. S. BUTTERS, N. SALMON, D. P.; JOHNSON, S. A. Comparison of the semantic networks in patients with dementia and amnesia. **Neuropsychology**, v. 9, n. 2, p. 177-186, 1995. Disponível em: <http://doi.apa.org/getdoi.cfm?doi=10.1037/0894-4105.9.2.177>. .

CHAN, A. S. BUTTERS, N. SALMON, D. P.; MCGUIRE, K. A. Dimensionality and clustering in the semantic network of patients with Alzheimer's disease. **Psychology and Aging**, v. 8, n. 3, p. 411-419, 1993. Disponível em: <http://doi.apa.org/getdoi.cfm?doi=10.1037/0882-7974.8.3.411>.

CHAN, A. S. SALMON, D. P. BUTTERS, N.; JOHNSON, S. A. Semantic network abnormality predicts rate of cognitive decline in patients with probable Alzheimer's disease. **Journal of the International Neuropsychological Society**, v. 1, p. 297-303, 1995. Disponível em: <http://www.journals.cambridge.org/abstract_S1355617700000291>.

CHAN, A. S. SALMON, D. P. NORDIN, S. MURPHY, C.; RAZANI, J. Abnormality of Semantic Network in Patients with Alzheimer's Disease: Evidence from Verbal, Perceptual, and Olfactory Domainsa. **Annals of the New York Academy of Sciences**, v. 855, n. 1, p. 681-685, 1998. Blackwell Publishing Ltd. Disponível em: <http://dx.doi.org/10.1111/j.1749-6632.1998.tb10645.x>.

CHARNIAK, E.; MCDERMOTT, D. **Introduction to artificial intelligence**. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1985.

COLLINS, A. M.; LOFTUS, E. F. A spreading-activation theory of semantic processing. **PsychologicalReview**, v. 82, p. 407–428, 1975.

CRIADO, R. FLORES, J. HERNÁNDEZ-BERMEJO, B. PELLO, J.; ROMANCE, M. Effective measurement of network vulnerability under random and intentional attacks.

Journal of Mathematical Modelling and Algorithms, v. 4, n. 3, p. 307-316, 2005. Disponível em: <http://www.springerlink.com/index/10.1007/s10852-005-9006-1>.

CSARDI, G.; NEPUSZ, T. The igraph software package for complex network research. **InterJournal**, v. Complex Sy, p. 1695, 2006. Disponível em: <http://igraph.sf.net>.

DEARHOLT, D. W.; SCHVANEVELDT, R. W. Properties of pathfinder networks. **Pathfinder associative networks**. p.1-30, 1990. Norwood, NJ, USA: Ablex Publishing Corp. Disponível em: <http://portal.acm.org/citation.cfm?id=119801.119802>.

DEESE, J. The Structure of Associations in Language and Thought. ,1965.

DOROGOVTSEV, S. N.; MENDES, J. F. F. Evolution of networks. **Advances in Physics**, v. 51, p. 1079-1187, 2002.

DUVIGNAU, KARINE; GAUME, B. Linguistic, Psycholinguistic and Computational approaches to the lexicon: For early verb-learning based on analogy. **Cognitive Systems**, , n. Special issue on learning, p. 255-269, 2004.

ERDÖS, P. A. R. On random graphs. **Publicationes Mathematicae**, v. 6, p. 290-297, 1959.

FERRER I CANCHO, R. SOLE, R. V.; KÖHLER, R. Patterns in syntactic dependency networks. **Physical Review E**, v. 69, p. 051915, 2004.

FREEMAN, L. Centrality in social networks conceptual clarification. **Social Networks**, v. 1, n. 3, p. 215-239, 1979. Disponível em: <http://dx.doi.org/10.1016/0378-8733(78)90021-7>.

GARRARD, P. LAMBON RALPH, M. A; PATTERSON, K. PRATT, K. H.; HODGES, J. R. Semantic feature knowledge and picture naming in dementia of Alzheimer’s type: a new approach. **Brain and language**, v. 93, n. 1, p. 79-94, 2005. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15766770>.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization and Machine Learning**. 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

GOLDSMITH, T. E.; DAVENPORT, D. M. Assessing structural similarity of graphs. **Pathfinder associative networks**. p.75-87, 1990. Norwood, NJ, USA: Ablex Publishing Corp. Disponível em: <http://portal.acm.org/citation.cfm?id=119801.119806>.

HALL, M. FRANK, E. HOLMES, G. et al. The WEKA data mining software: An update. **ACM SIGKDD Explorations Newsletter**, v. 11, n. 1, p. 10–18, 2009. ACM. Disponível em: <http://portal.acm.org/citation.cfm?id=1656274.1656278>. Acesso em: 2/2/2011.

HAUSER, M. D. CHOMSKY, N.; FITCH, W. T. The faculty of language: What is it, who has it, and how did it evolve. **Science**, v. 298, p. 569-1579, 2002.

HILLS, T. T. MAOUENE, M. MAOUENE, J. SHEYA, A.; SMITH, L. Longitudinal analysis of early semantic networks: preferential attachment or preferential acquisition? **Psychol Sci**, v. 20, n. 6, p. 729-739, 2009. Disponível em: <http://www.biomedsearch.com/nih/Longitudinal-analysis-early-semantic-networks/19470123.html>. .

HODGES, J. R. SALMON, D. P.; BUTTERS, N. Semantic memory impairment in Alzheimer’s disease: Failure of access or degraded knowledge. **Neuropsychologia**, v. 30, p. 301-314, 1992.

HOLTE, R. C. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. **Machine Learning**, v. 11, n. 1, p. 63-90, 1993. Hingham, MA, USA: Springer Netherlands. Disponível em: <http://dx.doi.org/10.1023/A:1022631118932>.

HOWARD, D.; PATTERSON, K. Pyramids and palm trees: A test of semantic access from pictures and words. Thames Valley Publishing. **Lambon Ralph, M. A., McClelland, J. L., Patterson, K., Galton, C. J., & Hodges, J. R**, v. 13, n. 3, p. 341-356, 1992.

JACOB, F. Evolution and tinkering. **Science**, v. 1964295, p. 1161-1166, 1977.

JOHN, G. H.; LANGLEY, P. Estimating Continuous Distributions in Bayesian Classifiers. p.338-345, 1995. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.8.3257>.

KE, J. Complex networks and human language. **CoRR**, v. abs/cs/070, 2007. Disponível em: <http://arxiv.org/abs/cs/0701135>.

KIM, M.; THOMPSON, C. K. Verb deficits in Alzheimer's disease and agrammatism: Implications for lexical organization. **Brain and Language**, v. 88, n. 1, p. 1-20, 2004. Disponível em: <http://www.sciencedirect.com/science/article/B6WC0-496FN4X-1/2/d547854da667c9290e72d80490aa566c>.

KOHAVI, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. IJCAI. **Anais** p.1137-1145, 1995. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.529>.

KOHAVI, R. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. Second International Conference on Knoledge Discovery and Data Mining. **Anais**, 1996. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.57.4952>.

KOLB, B.; WHISHAW, I. Q. Fundamentals of human neuropsychology. ,2003.

LAISNEY, M. GIFFARD, B. BELLIARD, S. et al. When the zebra loses its stripes: Semantic priming in early Alzheimer's disease and semantic dementia. **Cortex; a journal devoted to the study of the nervous system and behavior**, p. 1-12, 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/20089250>. Acesso em: 25/9/2010.

LAKOFF, G. Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. Chicago: University of Chicago Press. ,1987.

LANDWEHR, N. HALL, M.; FRANK, E. Logistic Model Trees. **Machine Learning**, v. 59, n. 1-2, p. 161-205, 2005.

MAC-KAY, A. P. M. G. ASSÊNCIO-FERREIRA, V.; FERRI-FERREIRA, T. **Afasias e demências avaliação e tratamento fonoaudiológico**. São Paulo, 2003.

MANSUR, L. L. CARTHERY, M. T. CARAMELLI, P.; NITRINI, R. Linguagem e cognição na doença de Alzheimer. **Psicologia: Reflexão e Crítica**, v. 18, n. 3, p. 300-307, 2005. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-79722005000300002&lng=pt&nrm=iso&tlng=pt>.

MATTIS, S. Mental status examination for organic mental syndrome in the elderly patient. In: R. Bellack; B. Karasu (Eds.); **Geriatric Psychiatry**. p.77-121, 1976. New York: Grune & Stratton.

MILLER, G. A. FELLBAUM, C. GROSS, D.; MILLER, K. J. Introduction to WordNet: An on-line lexical database. **International Journal of Lexicography**, , n. 3, p. 235-244, 1990.

MOTTER, A. E. MOURA, A. P. S. DE; LAI, Y. C.; DASGUPTA, P. Topology of the conceptual network of language. **Physical Review E**, v. 65, n. 6, p. 65102, 2002. APS. Disponível em: <http://link.aps.org/doi/10.1103/PhysRevE.65.065102>.

MÖLSÄ, P. K. MARTTILA, R. J.; RINNE, U. K. Survival and cause of death in Alzheimer's disease and multi-infarct dementia. **Acta Neurologica Scandinavica**, v. 74, n. 2, p. 103-107, 1986. Blackwell Publishing Ltd. Disponível em: <http://dx.doi.org/10.1111/j.1600-0404.1986.tb04634.x>.

NELSON, D. L. MCEVOY, C. L.; SCHREIBER, T. A. The University of South Florida word association norms. ,1999. Disponível em: <http://www.usf.edu/FreeAssociation>.
.

NELSON, D. L. MCKINNEY, V. M.; JANCZURA, G. A. Interpreting the Influence of Implicitly Activated Memories on Recall and Recognition. **Psychological Review**, v. 105, p. 299, 1998.

NEWMAN, M. E. J. The structure and function of complex networks. **SIAM Review**, v. 45, 2003. Disponível em: <http://arxiv.org/abs/cond-mat/0303516>.

NEWMAN, M. E. J. Analysis of weighted networks. **Physical Review E**, v. 70, p. 056131, 2004.

ORTIZ, K. Z. **Distúrbios Neurológicos Adquiridos: Linguagem e Cognição**. 2nd ed. Barueri, 2009.

PAGE, L. BRIN, S. MOTWANI, R.; WINOGRAD, T. The PageRank Citation Ranking: Bringing Order to the Web. ,1998. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1768>.

PARENTE, M. A. M. P. VILLAVICENCIO, A. SIQUEIRA, C. et al. Lexical Bootstrapping Hypothesis (LBH) and conventionality: The contribution of a crosslinguistic study in verb acquisition by Chinese Mandarin- and Brazilian Portuguese-speaking children. In: D. Bittner; N. Kuehn (Eds.); eds ed., 2011. Lexical Bootstrapping.

PERAITA, H. DÍAZ, C.; ANLLO-VENTO, L. Processing of semantic relations in normal aging and Alzheimer's disease. **Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists**, v. 23, n. 1, p. 33-46, 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/17961974>.

QUILLIAN, M. R. Semantic memory. **In Marvin Minsky**, 1968.

QUINLAN, R. J. **C4.5: programs for machine learning, San Mateo: Morgan Kaufmann**. Morgan Kaufmann Publishers Inc., 1993.

R DEVELOPMENT CORE TEAM. R: A Language and Environment for Statistical Computing. ,2010. Vienna, Austria. Disponível em: <http://www.r-project.org>.

ROGERS, S. L.; FRIEDMAN, R. B. The underlying mechanisms of semantic memory loss in Alzheimer's disease and semantic dementia. **Neuropsychologia**, v. 46, n. 1, p. 12-21, 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/17897685>.

ROGET, P. M. Roget's Thesaurus of English Words and Phrases. ,1911.

SANTOS, A. VALDEZ, G. G. VILLAVICENCIO, A.; SALLES, J. Investigating characteristics of semantic networks of verbs in patients with Alzheimer ' s disease. **Proceedings of Interdisciplinary Workshop on Verbs**, , n. The Identification and Representation of Verb Features, 2010.

SCHVANEVELDT, R. W. DURSO, F. T. GOLDSMITH, T. E. et al. Measuring the structure of expertise. **International journal of man-machine studies**, v. 23, n. 6, p. 699-728, 1985. Academic Press.

SCLIAR-CABRAL, L. Portuguese Florianopolis Corpus. **TalkBank**, 2004.

SHANNON, P. MARKIEL, A. OZIER, O. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. **Genome research**, v. 13, n. 11, p. 2498-2504, 2003. Institute for Systems Biology, Seattle, Washington 98103, USA. Disponível em: <http://dx.doi.org/10.1101/gr.1239303>.

SIGMAN, M.; CECCHI, G. A. Global organization of the Wordnet lexicon. **Proc. Nat. Acad. Sci**, v. 99, n. 3, p. 1742-1747, 2002.

SOLE, R. V. MURTRA, B. C. VALVERDE, S.; STEELS, L. Language Networks: their structure, function and evolution. **Trends in Cognitive Sciences**, , n. 22, p. 1-9, 2006. Citeseer. Disponível em: <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Language+Networks:+their+structure,+function+and+evolution#0>.

SOLÉ, R. V.; FERRER I CANCHO, R. The small world of human language. **Proceedings. Biological sciences / The Royal Society**, v. 268, n. 1482, p. 2261-5, 2001. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/11674874>.

STATISTICS, L. B.; BREIMAN, L. Random Forests. Machine Learning. **Anais** p.5-32, 2001.

STEYVERS, M.; TENENBAUM, J. B. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. **Cognitive Science**, v. 29, n. 1, p. 41-78, 2005.

TABERT, M. H. LIU, X. DOTY, R. L. et al. A 10-item smell identification scale related to risk for Alzheimer’s disease. **Annals of Neurology**, v. 58, n. 1, p. 155-160, 2005. Wiley Subscription Services, Inc., A Wiley Company. Disponível em: <http://dx.doi.org/10.1002/ana.20533>.

THOMPSON, C. K. Unaccusative verb production in agrammatic aphasia: the argument structure complexity hypothesis. **Journal of Neurolinguistics**, v. 16, n. 2-3, p. 151-167, 2003. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0911604402000143>.

THOMPSON, C. K.; SHAPIRO, L. P. Complexity in treatment of syntactic deficits. **American journal of speech-language pathology / American Speech-Language-Hearing Association**, v. 16, n. 1, p. 30-42, 2007. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2238729&tool=pmcentrez&rendertype=abstract>.

TONIETTO, L. VILLAVICENCIO, A. SIQUEIRA, M. et al. A especificidade semântica como fator determinante na aquisição de verbos. **PSICO**, v. 39, p. 343-351, 2008.

WATKINS, A.; BOGGESS, L. A New Classifier Based on Resource Limited Artificial Immune Systems. Evolutionary Computation. **Anais** p.1546-1551, 2002. Washington, DC, USA: IEEE Computer Society.

WATTS, D. J. **Small Worlds: the Dynamics of Networks between Order and Randomness**. 1999.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of small-world networks. **Nature**, v. 393, p. 440-442, 1998.

ZAIANE, O. R. Web Mining: Concepts, Practices and Research. In: SBC (Ed.); Simpósio Brasileiro de Banco de Dados. **Anais** v. 15, 2000. João Pessoa.

ZIPF, G. K. **Human Behavior and the Principle of Least Effort**. Addison-Wesley (Reading MA), 1949.