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

DANIEL CERATO GERMANN

Investigando a Influência de Fatores Lingüísticos na Organização Lexical de Verbos

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Prof. Dr. Luis Otavio Campos Alvares
Orientador

Porto Alegre, agosto de 2010.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Germann, Daniel Cerato

Investigando a Influência de Fatores Lingüísticos na Organização Lexical de Verbos / Daniel Cerato Germann – Porto Alegre: Programa de Pós-Graduação em Computação, 2010.

93 f.:il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2010. Orientador: Luis Otavio Campos Alvares.

1.Lingüística Computacional. 2.Grafos. I. Alvares, Luis Otavio Campos. II. Investigando a Influência de Fatores Lingüísticos na Organização Lexical de Verbos.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos Alexandre Netto
Vice-Reitor: Prof. Rui Vicente Oppermann
Pró-Reitor de Pós-Graduação: Prof. Aldo Bolten Lucion
Diretor do Instituto de Informática: Prof. Flávio Rech Wagner
Coordenador do PPGC: Prof. Álvaro Freitas Moreira
Bibliotecária-Chefe do Instituto de Informática: Beatriz Regina Bastos Haro

AGRADECIMENTOS

Agradeço a todos os que fizerem deste momento possível, em especial a meus pais, Adriana e José Carlos, e irmão Lucas, pelo apoio moral e psicológico, mesmo nos momentos mais difíceis. Agradeço também à Luciana, meu amor, cuja paciência e dedicação revelaram-se surpreendentes e cruciais para o alcance deste objetivo. Se existe vida após o mestrado, espero vivê-la contigo.

Agradeço à Profa. Dra. Aline Villavicencio, que iniciou este trabalho com dedicação e eficiência, e ao Prof. Dr. Luis Otavio Campos Alvares, que aceitou terminá-lo. Agradeço à Profa. Dra. Maity Simone Guerreiro Siquera, pelas conversas, orientações e conselhos, tanto no âmbito acadêmico quanto pessoal. À Profa. Dra. Maria Alice de Mattos Pimenta Parente, pela possibilidade de utilização de seus dados nesta pesquisa, bem como à Dra. Lauren Tonietto, que realizou a coleta destes. Agradeço aos bolsistas Bruno Menegola e a Gustavo Garcia Valdez, pela presteza e eficiência com que desenvolveram todas as demandas às quais lhes foram entregues. Agradeço ao NAE (Núcleo de Assessoria Estatística da UFRGS), por toda a dedicação demonstrada nesse auxílio estatístico, em especial à Profa. Elsa Mundstock, ao bolsista Luciano Guimarães e à estudante Silvana Schneider.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

G	Grafo
G1	Grafo do grupo 1
G2	Grafo do grupo 2
G3	Grafo do grupo 3
PolWord	Valores de polissemia extraídos do WordNetBR
PolHou	Valores de polissemia extraídos do dicionário eletrônico Houaiss
FreqYahoo	valores de freqüência de observação extraídos da API do buscador ‘Yahoo!’
FreqFlorian	Valores de freqüência de observação extraídos do corpus ‘Florianópolis’
CompSint	Valores de complexidade sintática
CombYahooWord	Valores da combinação entre FreqYahoo e PolWord
CombYahooHou	Valores da combinação entre FreqYahoo e PolHou
DP	Desvio Padrão
V	Grupo de vértices de um grafo
E	Grupo de arcos de um grafo
S	Quantidade de subgrafos de um grafo
C	Coeficiente de clusterização de um grafo
C/s	Coeficiente de clusterização adaptado de um grafo
L	Tamanho do caminho mínimo médio de um grafo
diam	Diâmetro de um grafo
<k>	Conectividade média de um grafo
D	Densidade de um grafo
J	Coeficiente de Jaccard de um grafo
ANOVA	*ANalisys Of VAriance* (análise da variância)
SPSS	*Statistical Package for the Social Sciences*

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Esta dissertação utiliza simulações computacionais visando investigar a influência de alguns fatores lingüísticos na organização lexical de verbos, analisando os processos de aquisição e uso. Os fatores testados são: freqüência de observação na linguagem, polissemia e complexidade sintática. Os dados utilizados foram obtidos por meio de tarefas psicolingüísticas de nomeação de ações, realizadas por crianças e adultos (falantes do Português brasileiro), posteriormente representados como grafos. Com base nos fatores lingüísticos, foram formuladas hipóteses relativas ao desenvolvimento da língua, testadas por meio de simulações computacionais denominadas 'involuções'. Os testes incluem métricas da teoria dos grafos e medidas de similaridade de conjuntos (coeficiente de Jaccard e suas componentes). Os resultados obtidos apontam para uma confirmação das hipóteses formuladas. Adicionalmente, permitiram verificar algumas características do desenvolvimento lingüístico, como o aumento do vocabulário e uma progressiva especialização.

Palavras-Chave: verbo, grafo, fator lingüístico, freqüência, polissemia, complexidade sintática, desenvolvimento lexical.

Investigating the Influence of Linguistic Factors in the Lexical Organization of Verbs

ABSTRACT

This dissertation uses computational simulations designed to investigate the influence of three linguistic factors in the lexical organization of verbs, analyzing the process of acquisition and use. The tested factors are: frequency of observation in the language, polysemy and syntactic complexity. The data used were obtained from psycholinguistic action naming tasks performed by children and adults (speakers of Brazilian Portuguese), and subsequently represented as graphs. Based on linguistic factors, hypotheses were formulated concerning the development of language, tested through simulations called 'involutions'. Tests include graph theory metrics and set similarity measures (Jaccard's coefficient and its components). Results suggest a confirmation of the given hypotheses. Additionally, allowed verification of some language development features, such as vocabulary growth and a progressive specialization.

Keywords: verb, graph, linguistic factor, frequency, polysemy, syntactic complexity, lexical development.

1 INTRODUÇÃO

Abordagens computacionais no campo lingüístico deram origem a tecnologias importantes, como tradução automática, a extração de informações a partir de fontes de dados (como a Internet) e a sumarização de textos. Entretanto, para que estas e outras facilidades evoluam, é preciso estudar a fundo os mecanismos envolvidos na comunicação. Somente entendendo melhor as estruturas mentais e os processos envolvidos na comunicação, será possível aprender de fato como funcionam a compreensão, representação e produção de informações – e então replicá-las artificialmente com maior fidelidade.

Estudos computacionais acerca da aquisição e representação da linguagem já foram realizados por diversos autores, comumente focando em algum aspecto mais específico desses assuntos. A aquisição baseada em similaridades sintáticas, por exemplo, foi alvo do trabalho de Yu (2006), enquanto Fazly et al. (2008) simularam o relacionamento estatístico entre palavras e componentes semânticas e Yu (2005) abordou a aquisição lexical com base em padrões visuais. Trabalhos como estes procuram elucidar os critérios que motivam a aquisição de conceitos, criando modelos computacionais que visam a replicação de resultados empíricos anteriores. Seguindo a mesma abordagem, mas com foco ligeiramente diferente, Xu e Tenenbaum (2007), simulam a generalização de conceitos e o aprendizado em diferentes níveis taxonômicos. Em outras palavras o objeto de estudo passa a ser a organização mental dos conceitos.

Complementando os modelos computacionais de aquisição e organização, estão os estudos sobre a estrutura da linguagem. Neste caso, corpora de palavras são analisados topologicamente, buscando similaridades (STEYVERS & TENENBAUM, 2005; GORMAN & CURRAN, 2007) que caracterizem as linguagens. Dessa forma, o foco passa da aquisição para a replicação de estruturas com as mesmas características que as linguagens reais. A existência de estudos lingüístico-computacionais tão diversos evidencia o quão amplamente estudada tem sido a linguagem em seus mais diversos aspectos. Entretanto, é possível perceber que o processo evolutivo, o modo como a utilização da linguagem (e as estruturas relacionadas) desenvolve-se no ser humano, recebeu pouca atenção até o momento.

Compreender o processo evolutivo da linguagem é essencial para um entendimento adequado e completo da aquisição. Afinal, os aspectos que determinam a aquisição e a organização dos conceitos podem mudar conforme o indivíduo envelhece. A influência da idade foi estudada por Coronges et al. (2007), que realizaram um estudo topológico comparativo entre dois grupos etários, mas analisando apenas as diferenças estruturais, sem investigar a influência de fatores específicos no processo. A presente pesquisa visa, portanto, realizar essa investigação, propondo e testando uma abordagem computacional para a análise direta da evolução lexical (com ênfase no estudo dos fatores que a influenciam).

Além de investigar o processo de evolução da linguagem, focaremos no estudo de verbos. O interesse nessa classe gramatical se justifica por sua importância comunicativa: verbos denotam eventos, ligando e relacionando conceitos, e possibilitando a construção de informações complexas, essenciais para a comunicação. Além disso, historicamente, eles foram alvo de poucos estudos (em especial quando comparados aos substantivos, a classe gramatical mais abordada), evidenciando, desta forma, um campo a ser explorado. A provável origem de uma quantidade significativamente menor de trabalhos focando em verbos é sua estrutura: não são objetos tão fechados quanto os substantivos, que possam ser diretamente analisados. Em vez disso, representam relações entre objetos, sendo, portanto, inerentemente mais abstratos (LAAKSO & SMITH, 2007; mas ver BLACK & CHIAT, 2003).

A escolha dos fatores lingüísticos a serem testados procurou abordar os aspectos mais diversos possíveis. O primeiro deles é a freqüência de observação do verbo dentro da produção lingüística. Uma vez que o aprendizado parece ser baseado em co-ocorrências de diversos fatores (como sugerido, entre outros, por YU, 2005; YU, 2006; XU & TENENBAUM, 2007; FAZLY et al., 2008), é natural o interesse em verificar se a quantidade de exposições a determinada palavra possui impacto na sua evolução lingüística. O segundo fator é a polissemia, que procura verificar o quão difícil é o aprendizado de um verbo por conta de seu conjunto de significados. O último parâmetro é a complexidade sintática, que procura verificar a dificuldade de aprendizado de acordo com a facilidade em utilizá-lo em construções lingüísticas (mais especificamente, a quantidade de objetos necessária para sua utilização).

A presente pesquisa baseia-se em hipóteses objetivas: supomos que os verbos com maior freqüência, maior polissemia e menor complexidade sintática possuem maior chance de serem aprendidos primeiro. Para verificar as hipóteses, foram construídos os vocabulários de três grupos de indivíduos, agrupados com base na idade, expressos por meio de grafos. Esses grafos foram iterativamente modificados: os verbos com maior freqüência, maior polissemia e menor complexidade sintática foram progressivamente eliminados[1], e o grafo alterado foi sendo comparado com o do grupo imediatamente mais novo. Essas comparações foram baseadas em métricas da teoria dos grafos e da teoria dos conjuntos, o que permitiu uma mensuração direta do impacto exercido por cada fator na evolução do léxico-mental verbal. Os resultados parecem comprovar, de fato, uma preferência dos indivíduos mais jovens pela utilização de verbos mais freqüentes, polissêmicos e sintaticamente simples.

Além da comprovação das hipóteses propostas, a presente pesquisa inova ao simular a evolução de uma forma direta (propondo o algoritmo *network involution*, apresentado no capítulo 3). Também inova ao incluir métricas da teoria dos conjuntos (apresentadas no capítulo 2) na análise realizada.

A dissertação está organizada conforme descrito a seguir. O capítulo 2 contextualiza a pesquisa em relação à metodologia utilizada, aos objetivos e à escolha dos dados. O capítulo 3 apresenta o modelo de representação, bem como a dinâmica dos testes executados. O capítulo 4 apresenta e explica os resultados encontrados, e o capítulo 5 sumariza as conclusões, além de sugerir trabalhos futuros.

[1] Inicialmente, cada fator foi analisado separadamente. Posteriormente, duas combinações de fatores foram testadas.

2 BASES TEÓRICAS

Neste capítulo, serão apresentadas as bases teóricas que fundamentam a pesquisa realizada no que tange a aquisição da linguagem, o desenvolvimento humano e abordagens computacionais. Em especial, serão abordadas as métricas utilizadas na avaliação dos resultados e a motivação para a escolha dos fatores lingüísticos testados. Além disso, será apresentado um panorama de outros trabalhos que se relacionam ao tema em estudo, de modo a contextualizar a metodologia empregada e os objetivos definidos.

2.1 Contextualizando a Pesquisa

Segundo Goldberg (1999), o aprendizado de verbos parece dar-se de forma incremental. Verbos, e também estruturas sintáticas (o modo como um verbo pode ser usado, em relação ao sujeito, objetos, etc.), são aprendidos, inicialmente, um a um. O domínio inicial dá-se pela imitação e, posteriormente, avança por meio da generalização do conhecimento sintático: as estruturas sintáticas adquiridas são aplicadas sobre os verbos conhecidos e reação obtida é verificada, de modo a descobrir se a aplicação foi correta ou não. Conforme mais verbos vão sendo aprendidos e testados, o vocabulário cresce e a linguagem desenvolve-se.

No contexto da evolução lexical, a presente pesquisa procura elucidar como esse processo de aprendizagem se dá. Os fatores envolvidos são compreendidos mais facilmente quando avaliamos a própria evolução humana. A criança nasce com um aparato biológico complexo, bastante similar ao dos adultos – ao menos no que tange a percepção, uma vez que os sensores são os mesmos (visão, audição, tato, etc.). Entretanto, a interpretação da informação dá-se de forma diferente, resultando em experimentações distintas das mesmas observações. Podemos citar ao menos duas diferenças importantes: falta de conhecimento de mundo (cujo acúmulo acaba sendo útil como guia na percepção e no processo de organização da informação) e limitações dos mecanismos de interpretação da informação percebida.

Estas diferenças podem ser comprovadas na nomeação de objetos. Segundo Mervis (1984), por desconhecer a importância de algumas características, a criança podia acabar chamando um cofrinho redondo de 'bola', em vez de 'cofre' por exemplo. Neste caso, a criança acaba valorizando características de forma em detrimento de funcionalidade (a possibilidade de ‘armazenar dinheiro’). Mandler (2003) e Younger (2003), por sua vez, sugerem que esta valorização da forma na infância é parte do processo natural de aprendizado, e que a ‘funcionalidade’ passa a ser percebida mais tarde no desenvolvimento cognitivo.

Younger (2003) defende ainda que as crianças mais novas seriam capazes de manter apenas uma categoria conceitual[2] em mente por vez, classificando os objetos a sua volta como pertencentes a ela ou não. Por exemplo, se o infante está analisando, em determinado momento, animais terrestres, ele provavelmente agruparia cachorros, gatos e vacas como pertencentes à categoria, e pássaros, bolas, portas e prédios como não pertencentes. Segundo Mandler (2003), essa limitação motiva a criação de categorias mais amplas, i.e., com poucas características associadas. Afinal, uma categoria mental com poucas características permite dividir melhor[3] os objetos do mundo. Dessa forma, experiências associadas a um objeto podem servir como base para o julgamento de experiências com muitos outros objetos (por conta de pertencerem ou não à mesma categoria mental). Conforme o indivíduo amadurece, torna-se possível analisar múltiplas categorias mentais ao mesmo tempo.

O conteúdo dessas categorias iniciais está diretamente associado a habilidades natas de reconhecimento humano e correlação de sinais físicos. Por exemplo, as crianças possuem a habilidade natural de perceber padrões auditivos e visuais (MARCUS et. al, 1999), buscando, constantemente, identificar padrões dentro do seu conjunto de percepções. Dessa forma, ainda que não consigam compreender o mundo a sua volta, elas parecem procurar por referenciais sobre os quais focar a atenção. Além disso, elas naturalmente reconhecem vozes (MEHLER, 1985) e rostos (SPELKE, 1985) humanos, o que sugere um aprendizado otimizado: o comportamento de outros humanos deve ser similar ao comportamento que o próprio ser deve apresentar, sendo, portanto, uma boa base de observação. Por fim, as crianças possuem uma propensão natural à imitação (MELTZOFF & MOORE, 1985; GOLDBERG, 1999), o que poderia fazer com que aprendessem mais rapidamente.

Em suma, o modo como ocorre o desenvolvimento infantil favorece compreensões de mundo distintas de acordo com a idade, distinções estas que se refletem também no desenvolvimento lingüístico. Sabendo que a criança possui uma compreensão parcial (limitada por suas capacidades mentais ainda em desenvolvimento) e que aprende de forma progressiva, o objetivo da presente pesquisa é verificar quais fatores favorecem o aprendizado de determinados verbos em detrimento de outros. Entre os fatores – lingüísticos e psicolingüísticos – preponderantes, destacam-se idade de aquisição (ELLIS & MORRISON, 1998; ELLIS & RALPH, 2000) e freqüência de observação (MORRISON & ELLIS, 1995), além de características semânticas (BREEDIN et. al, 1998; BARDE et al., 2006) e sintáticas (GOLDBERG, 1999; THOMPSON et. al, 2003; FERRER-I-CANCHO et al., 2004) das palavras. Considerando a importância destes fatores (apresentados na seção 2.2) alguns deles serão testados em simulações computacionais de modo a verificar a sua influência no processo evolutivo do léxico mental verbal.

---

[2] 'Categorias conceituais' podem ser descritas como definições de conceitos; são agrupamentos de características. Os objetos do mundo podem ser classificados como pertencentes ou não a ela e em diversos graus (pertencendo muito ou pouco, por exemplo). Para mais detalhes sobre conceitos e categorias, consulte Lakoff (1987), Evans e Green (2006) e Schmid e Ungerer (2006).

[3] Uma categoria mental com muitas características associadas, como 'carro de corrida vermelho' teria poucos objetos considerados como 'pertencentes' e muitos como 'não pertencentes'. Uma categoria com menos características, como 'objeto que se move', divide melhor os objetos do mundo: a quantidade de objetos 'pertencentes' aproxima-se da de 'não-pertencentes'.

2.1.1 Estudos Computacionais

Nos últimos anos, tem havido um crescente interesse na investigação da aquisição da linguagem utilizando modelos computacionais. Por exemplo, alguns trabalhos têm investigado as propriedades da linguagem, como o *fast-mapping*[4] e o *age-of-acquisition effect*[5] (ELLIS & RALPH, 2000; LI et al., 2004; REGIER, 2005; HORST et al., 2006). Outros têm simulado o aprendizado (SISKIND, 1996; YU, 2005; YU, 2006; FAZLY et al., 2008; PARISIEN & STEVENSON, 2009) e suas particularidades como a generalização de conceitos (XU & TENENBAUM, 2007) e a descoberta de palavras com multiplicidade de significado (DOROW & WIDDOWS, 2003), além do desenvolvimento lexical (STEYVERS & TENENBAUM, 2005; GORMAN & CURRAN, 2007) e da plasticidade cerebral (PLAUT, 1996). A evolução da linguagem dentro da sociedade também foi investigada (MEHLER, 2007; GONG et al., 2008).

Estudos computacionais focando em falantes do Português Brasileiro são poucos. Dentre estes, destacamos o trabalho de Soares et al. (2005), que reporta um estudo fonético, e também de Antiqueira et al. (2007), que correlaciona métricas de teoria dos grafos e métricas subjetivas de qualidade de texto. Tonietto et al. (2008) e Tonietto (2009) analisam a influência de aspectos pragmáticos, como a convencionalidade de uso, sobre a organização lexical dos verbos, e observam que os adultos tendem a preferir palavras mais convencionais do que crianças. Os dados obtidos nestes dois trabalhos serviram de base para a construção dos grafos utilizados nas simulações computacionais desta pesquisa. Tonietto (2009) realiza, inclusive, um estudo similar, construindo grafos para expressar o léxico mental, e realizando análises topológicas e estatísticas. A presente pesquisa diverge, entretanto, no método de construção dos grafos, nos fatores analisados, e na utilização de involuções (explicados no capítulo 3)

Na presente pesquisa, os grafos expressam vocabulários de grupos de indivíduos, sendo que cada vértice representa um verbo, e cada ligação representa um compartilhamento de significado (como será visto em mais detalhes no capítulo 3). Os significados são, portanto, indiretamente considerados, uma vez que se encontram implícitos nas arestas. Dentre os trabalhos que utilizaram teoria dos grafos no âmbito semântico, destacamos Sinha e Mihalcea (2007) e Navigli e Lapata (2007), que utilizaram estas estruturas na desambiguação de palavras.

Medidas de teoria dos grafos também podem ser utilizadas na comparação de sistemas complexos, como as linguagens. Ferrer-i-Cancho et al. (2004) realizaram a análise de linguagens européias, enquanto Motter et al. (2002) analisaram o inglês. As linguagens humanas em geral foram caracterizadas de acordo com métricas da teoria dos grafos por Masucci e Rodgers (2006). Gaume et al. (2006), por sua vez, também utilizaram estes recursos no estudo da evolução do significado nas linguagens, traçando um paralelo com a dinâmica de uma sociedade. No mesmo sentido, Sigman e Cecchi (2002) utilizam métricas da teoria dos grafos para extensivamente analisar propriedades da WordNet[6], em especial o impacto da polissemia sobre ela. De forma similar, Gorman

[4] '*Fast-mapping*' refere-se à habilidade das crianças de rapidamente aprenderem o significado de uma palavra em poucas tentativas.

[5] '*Age-of-acquisition effect*', refere-se a propriedades especiais percebidas em palavras aprendidas mais cedo, como rápido reconhecimento e lembrança rápida. Ou seja, é uma relação existente entre desempenho e idade de aquisição dos conceitos.

[6] WordNet é um importante banco de dados que relaciona e organiza palavras segundo propriedades semântico-conceituais (como a definição) e lexicais (como a classe gramatical).

e Curran (2007) utilizam métricas de teoria dos grafos para analisar três redes semânticas lingüísticas (entre elas a WordNet), depois de reestruturá-las utilizando a sinonímia e a homonímia (separadamente) como a ligação entre as palavras. Steyvers e Tenenbaum (2005) realizaram um trabalho similar, investigando as propriedades estruturais (estatísticas e organizacionais) de redes semânticas (entre elas a WordNet). Dentre as análises realizadas, os autores relacionam propriedades lingüísticas (em especial a freqüência e a idade de aquisição) com propriedades dos grafos (em especial o grau do vértice). Adicionalmente, eles propõem um modelo – gerativista – de crescimento que procura apresentar as mesmas características encontradas na etapa de análise. Utilizando a inserção iterativa de nodos, esse modelo demonstra que, por meio de regras simples de associação entre os vértices (quanto maior o número de conexões de um vértice existente, maior a probabilidade de ele receber uma conexão do vértice sendo inserido), é possível construir uma estrutura bastante similar a outras que expressam a representação humana. Em outras palavras, o modelo simula a evolução lingüística humana.

Da mesma forma que Steyvers e Tenenbaum (2005), De Deyne e Storms (2008) realizaram a análise de propriedades lingüísticas utilizando teoria dos grafos. Os autores encontraram uma associação entre a centralidade e a idade de aquisição: vértices centrais tendem a ser adquiridos mais cedo e a apresentarem grande freqüência. Uma outra abordagem para o estudo de propriedades lingüísticas é apresentado por Coronges et al. (2007). Em vez de utilizar recursos produzidos artificialmente (como a WordNet), os autores criaram grafos representando a estrutura cognitiva de grupos de indivíduos separados por idade e analisaram segundo diversas métricas de teoria dos grafos. Os grupos puderam então ser comparados de acordo com as métricas encontradas.

A abordagem proposta no presente trabalho segue Steyvers e Tenenbaum (2005) no sentido de iterativamente modificar um grafo, mas difere em método: utilizamos ‘involuções’ em vez de ‘evoluções’, como será explicado no capítulo 3. Difere também em objetivo: nossas modificações são motivadas pelo estudo do impacto de fatores lingüísticos no processo evolutivo, não pela produção de uma topologia com uma organização específica. Seguimos também De Deyne e Storms (2008), no sentido de relacionar diretamente fatores lingüísticos e teoria dos grafos, além de Coronges et al. (2007), no sentido de comparar redes de diferentes grupos etários com estas métricas.

2.2 Fatores de Influência na Aquisição Lexical Verbal

Como mencionado na seção 2.1, a presente pesquisa visa investigar o impacto de fatores lingüísticos na evolução lexical. Dentre os vários existentes, três foram escolhidos para serem testados: freqüência de observação, polissemia e complexidade sintática. A escolha levou em consideração as restrições cognitivas infantis mencionadas no início deste capítulo: o pouco (ou nenhum) conhecimento de mundo e a capacidade mental restrita. Na presente pesquisa, assumimos que verbos com maior freqüência de observação, maior polissemia e menor complexidade sintática tendem a ser adquiridos mais cedo. Como será apresentado a seguir, verbos que estão de acordo com estes princípios tendem a apresentar maior utilidade, além de serem mais facilmente compreendidos, portanto facilitando a aquisição.

2.2.1 Freqüência de Observação

A importância da freqüência de observação foi comprovada por Howes e Solomon (1951) e Solomon e Postman (1952): palavras percebidas uma quantidade maior de

vezes são mais facilmente lembradas (tanto palavras ‘com’ quanto ‘sem’ sentido). A importância da freqüência também é apontada por Goldberg (1999): os verbos com maior freqüência de *input* (para as crianças) são os com utilização mais fiel ao significado original, além de serem aprendidos primeiro.

A freqüência de observação também se relaciona diretamente a outros fatores importantes, como a semântica: verbos com significados mais gerais e maior polissemia tendem a ser mais freqüentemente empregados. Isso foi verificado tanto em pessoas normais (GOLDBERG, 1999; BARDE et al., 2006) quanto em pessoas com problemas lingüísticos (BREEDIN et al., 1998; mas veja KIM & THOMPSON, 2004), sendo diretamente relacionado à utilidade: verbos mais gerais podem ser utilizados numa maior quantidade de contextos, portanto tendem a aparecer mais na linguagem.

Por outro lado, a importância da freqüência foi contestada por alguns autores. Foi demonstrado que algumas medidas de desempenho anteriormente associadas a ela, como grande velocidade de nomeação (MORRISON & ELLIS, 1995) e grande assertividade na nomeação de figuras (DAVIDOFF & MASTERSON, 1996) estavam associadas, na verdade, à idade de aquisição. O engano deu-se por conta da grande correlação entre os dois fatores (MORRISON & ELLIS, 1995; ELLIS & MORRISON, 1998; ELLIS & RALPH, 2000; LI et al., 2004), variando entre 0,40 e 0,71 (diversos estudos em MORRISON & ELLIS, 1995). A conclusão central destes estudos é que a freqüência é um fator importante para o desempenho de diversas tarefas lingüísticas, mas o é principalmente por causa de sua influência na idade de aquisição (palavras com maior freqüência tendem a ser adquiridas mais cedo). Quando a idade de aquisição é controlada, alguns dos benefícios associados à freqüência desaparecem.

Nesse contexto, a presente pesquisa diverge em foco quando comparada a outros trabalhos: enquanto outros estudaram a influência da freqüência e da idade de aquisição no desempenho de tarefas lingüísticas, nós focamos na influência da freqüência (e outros parâmetros) na idade de uso. O objetivo é demonstrar que palavras com maior freqüência de observação tendem a ser dominadas primeiro e utilizadas mais cedo. Essa verificação dar-se-á por meio da análise do vocabulário de diferentes grupos etários e das simulações computacionais. Ainda que idade de aquisição e idade de uso possam não coincidir perfeitamente, utilizaremos esta como indicação da ocorrência daquela. Afinal, a idade de aquisição não pode ser medida com exatidão (uma vez que não existe método objetivo para determinar quando o indivíduo aprendeu a palavra de fato), demandando sempre alguma forma de aproximação.

2.2.2 Polissemia

As dificuldades semânticas de aquisição são intrínsecas à palavra; dizem respeito à quão custosa será a interpretação e a utilização da palavra apenas por conta do significado. Existem diversos fatores semânticos influenciando nessa dificuldade, todos inter-relacionando-se, como será visto a seguir. Dentre eles, a polissemia foi escolhida como o foco da presente pesquisa não apenas por sua importância, mas por permitir solucionar uma série de problemas inerentemente relacionados a fatores semânticos (como também será visto a seguir). A polissemia refere-se à capacidade de uma palavra apresentar múltiplos significados: o verbo “picar”, por exemplo, pode significar ‘furar’, ‘reduzir algo a pequenos pedaços’ ou ‘causar ferimento’, entre outras possibilidades.

Considerando-se os fatores semânticos, a ‘complexidade semântica’, uma medida relativa de complexidade, é um dos mais importantes. Assumindo-se que os conceitos são definidos por meio da composição de outros mais simples (por exemplo, ‘correr’

pode ser definido como ‘ir de modo rápido’), um verbo é dito ‘mais complexo’ do que outro se aquele possuir uma quantidade maior de componentes semânticas do que este[7] (BREEDIN et al., 1998). A generalidade (BREEDIN et al., 1998) é uma medida similar à complexidade semântica, mas que desconsidera a estrutura de definição, classificando subjetivamente as palavras em diferentes amplitudes (muito ou pouco gerais). O peso semântico (*semantic weight*; BREEDIN et al., 1998; GOLDBERG, 1999; BARDE et al., 2006), é uma medida similar, mas absoluta: um verbo é dito ‘leve’ (*light* verb) se não contiver nenhum outro verbo em sua definição (senão, ‘pesado’). Por fim, temos a polissemia, que mede a quantidade de significados diferentes de uma palavra. Os três primeiros conceitos (complexidade semântica, generalidade e peso semântico) estão diretamente relacionados, um como parte da definição do outro (BREEDIN et al., 1998; GOLDBERG, 1999; KIM & THOMPSON, 2004; BARDE et al., 2006). Afinal, palavras com menor complexidade, por serem mais simples, podem ser associadas a uma quantidade maior de contextos (KIM & THOMPSON, 2004; BARDE et al., 2006), o que faz com que acabem podendo assumir uma maior quantidade de significados, ou seja, maior polissemia. Por conta dessa flexibilidade no emprego, estes verbos também costumam ser mais freqüentes (GOLDBERG, 1999).

O estudo dos *light verbs* (verbos classificados como ‘leves’ em relação ao peso semântico), em especial, corrobora e expande as relações mencionadas entre os diversos fatores semânticos (GOLDBERG, 1999). Estes verbos estão associados aos seguintes fatores: idade de aquisição (estão entre os primeiros a serem adquiridos), generalidade (são mais gerais, portanto comumente mais polissêmicos) e freqüência (são mais freqüentes na língua), além de serem mais facilmente empregados por afásicos (BREEDIN et al., 1998; THOMPSON, 2003; THOMPSON et al. 2003; BARDE et al. 2006; mas veja KIM & THOMPSON, 2004), sugerindo maior importância para a cognição humana. Ainda que exista uma correlação entre polissemia e freqüência, o fato de esta não prever perfeitamente a idade de aquisição motiva um estudo mais aprofundado daquela (PARISIEN & STEVENSON, 2009).

Quanto à influência dos fatores semânticos na aquisição, Goldberg (1999) apresenta evidências de que os *light verbs* são adquiridos mais cedo, e que servem como base para o aprendizado de outros verbos. Uziel-Karl (2001), analogamente, demonstra que crianças tendem a preferir verbos mais gerais e polissêmicos (uma vez que, segundo o autor, maior generalidade está associada a maior polissemia).

Medir a dificuldade semântica não é uma tarefa simples, encontrando ao menos três desafios principais: (1) inexistência de um método imediato e objetivo para se quantificar a dificuldade de compreensão por conta da semântica, (2) dificuldade em se lidar com os múltiplos significados associados a cada palavra e (3) dificuldade em se delimitar perfeitamente o sentido de uma palavra, por conta de variações. Será mostrado que, por meio da medição do número de significados (i.e., contabilização da polissemia), é possível superar todos estes impedimentos potenciais.

O primeiro problema relaciona-se à dificuldade em se quantificar um aspecto subjetivo: ‘Como mensurar a dificuldade de compreensão de uma palavra?’ Uma solução computacional comum para se lidar com um valor de difícil obtenção é a

[7] Esta definição não é um consenso. Uma definição alternativa de ‘complexidade semântica’ foi apresentada por Kiran e Thompson (2003), baseando-se na tipicalidade: quanto mais típica a palavra, mais semanticamente simples ela é. Este uso é baseado no trabalho de Rosch (1975), cujos estudos focavam na similaridade entre uma palavra e sua categoria mais geral para definir a tipicalidade desta.

definição de uma heurística que se correlacione adeqüadamente com o que se está querendo medir. Os quatro fatores semânticos mencionados podem servir como heurística, mas eles não necessariamente trarão os mesmos resultados. De qualquer forma, uma vez que não existe uma forma objetiva de se medir a dificuldade compreensão, é difícil determinar a qualidade de cada um desses fatores.

O segundo problema relaciona-se diretamente à polissemia. É sabido que a maior parte das palavras possui múltiplos significados associados. Assim, se queremos determinar a dificuldade de compreensão de uma palavra, estes múltiplos significados devem ser considerados de alguma forma. Uma solução possível é assumir um significado como mais representativo e, desta forma, tratá-lo como o 'oficial'. Não será possível determinar a generalidade, complexidade ou peso semântico sem que essa escolha seja feita. Entretanto, as informações descartadas podem ser demasiadamente importantes, uma vez que nem toda palavra possui, de fato, um significado indiscutivelmente mais importante (o verbo 'fazer', por exemplo, dentre seus vários significados, pode representar 'executar alguma ação', 'fabricar alguma coisa', 'compor', 'provocar um efeito', 'exercer ou praticar uma atividade', etc.). Outra alternativa é apenas contabilizar todos estes significados, como será feito na presente pesquisa, e usar o resultado para medir a dificuldade compreensão, evita a necessidade da escolha.

O terceiro problema relaciona-se à ausência de limites bem definidos nos significados de um termo e à própria mutabilidade da língua: o que hoje é considerado uma gíria ou metáfora pode, um dia, ser incluído como significado literal. Como conseqüência, pessoas diferentes podem divergir acerca da extensão do significado de determinada palavra, ou mesmo na quantidade de significados existentes. Para evitar este problema (ainda que não o resolvendo), foram buscadas fontes confiáveis de informação, que foram tomadas como consenso teórico (o banco de dados WordNetBR e o dicionário Houaiss; ver capítulo 3 para maiores detalhes). Dessa forma, mesmo que outras fontes apresentem uma quantidade diferente de significados para uma palavra, ou que a extensão de um significado seja mais ou menos ampla, estas diferenças serão tratadas como variações normais e, desta forma, não levadas em consideração. Essa variação poderia acabar sendo significativa caso utilizássemos algum outro fator semântico (onde o significado em si teria de ser avaliado). Entretanto, uma vez que iremos apenas contabilizar a quantidade de significados diferentes, essa discrepância potencial tende a ser minimizada (pois a variação dessa quantidade tende a não ser grande).

Em suma, a simples existência da polissemia, bem como da mutabilidade lingüística, dificulta a utilização dos outros fatores semânticos, motivando sua utilização como heurística. Além disso, o problema da subjetividade nas definições das palavras acaba tornando-se secundário uma vez que o significado em si não importa (apenas a quantidade, e esta tende a variar pouco).

2.2.3 Complexidade Sintática

Como visto anteriormente, verbos expressam relações, em geral demandando outros referenciais para sua compreensão (BLACK & CHIAT, 2003). Maior número de referenciais significa maior número de focos de atenção necessários para que a compreensão da frase seja possível (para outros fatores que influenciam na lembrança de verbos, veja LAUDANNA et. al, 2004). A forma mais imediata de se expressar a multiplicidade de focos de atenção de um verbo é por meio de sua complexidade

sintática (DAVIDOFF & MASTERSON, 1996; THOMPSON, 2003; KIM & THOMPSON, 2004; LAAKSO & SMITH, 2007): o número e tipo de objetos gramaticais associados ao verbo. No Português, os verbos podem ser classificados progressivamente (em termos de complexidade sintática) como:

- Intransitivos: não demandam objetos. Exemplo: 'Choveu ontem.'
- Transitivos diretos: demandam um objeto direto, sem preposição. Exemplo: 'Paguei a conta. '
- Transitivos indiretos: demandam um objeto indireto, com preposição. Exemplo: 'Paguei à atendente. '
- Bitransitivos: demandam um objeto direto e outro indireto. Exemplo: 'Paguei a conta à atendente. '

Pelos exemplos fica claro que um mesmo verbo pode ser utilizado de várias formas, tendo, portanto, múltiplas complexidades sintáticas associadas. Na presente pesquisa, a complexidade associada a determinado verbo será a forma mais comum encontrada em um *corpus*.

A motivação para o estudo da sintaxe deve-se a seu papel fundamental na comunicação. A sintaxe é resultado da necessidade de se comunicar uma quantidade muito grande de informações (SCHOENEMANN, 1999), ou seja, são as regras sintáticas que permitem a qualidade e a complexidade da comunicação atual. Em outras palavras, foram necessidades semânticas (de comunicar informações mais complexas) que motivaram a existência da sintaxe. Além disso, existem regularidades sintáticas entre as linguagens, fruto de necessidades similares de comunicação. Estudar o impacto da sintaxe na aquisição da linguagem, portanto, é compreender melhor a dinâmica da transmissão de informações.

Schoenemann (1999) defende a idéia de que a sintaxe só existe para auxiliar na troca de informações semânticas, mas o relacionamento entre os dois fatores vai além. A dificuldade no aprendizado de um verbo, por exemplo, depende, entre outros fatores, de quais substantivos foram utilizados como sujeitos e objetos nas frases em que ele aparece (LAAKSO & SMITH, 2007). Davidoff e Masterson (1996) reforçam a idéia da relação entre forma e conteúdo nas frases, apontando que verbos transitivos, por terem maior quantidade de argumentos, são mais difíceis de serem lembrados, mas mais fáceis de serem identificados em material pictórico. Uma vez que estes argumentos referem-se a objetos reais, eles servem como âncoras na identificação do verbo. Barde et. al (2006) confirmaram a hipótese de que essa relação é ainda mais profunda: verbos semanticamente leves (*light verbs*), tendem a depender mais da sintaxe e menos da semântica na identificação, e o inverso para verbos pesados. Por possuírem menos características semânticas, são menos previsíveis a partir da observação semântica, delegando à análise sintática o processo de identificação. Se existe uma lesão na parte sintática do cérebro, os verbos mais leves serão mais afetados do que os pesados. Esse relacionamento entre lesões em porções cerebrais e deficiências relacionadas a tipos de verbos está em consonância com a existência de um decréscimo na utilização de verbos leves em um grupo com afasia agramática, mas não em um grupo com agramatismo. De fato, diversos estudos têm mostrado que indivíduos com déficits lingüísticos possuem dificuldades especialmente relacionadas à complexidade sintática (KIM & THOMPSON, 2000; THOMPSON, 2003; THOMPSON et al., 2003; KIM & THOMPSON, 2004; BARDE et. al, 2006). Em outras palavras, ainda que a sintaxe

possa ter sido criada principalmente para facilitar a comunicação (SCHOENEMANN, 1999), o relacionamento direto entre partes do cérebro e o processamento sintático é evidente.

Em suma, a complexidade sintática estará exprimindo a dificuldade extrínseca encontrada pelo indivíduo para compreender determinado verbo: o quão difícil é a cognição por conta dos relacionamentos necessários para a compreensão da palavra. A importância do fator deve-se a seu papel na comunicação, bem como à existência de relacionamentos com a semântica e o processamento biológico. Além disso, o fato de estarmos medindo a quantidade de focos de atenção é interessante, uma vez que crianças menores não dispõem de uma grande quantidade de recursos cognitivos (sendo, portanto, um fator preponderante para o aprendizado).

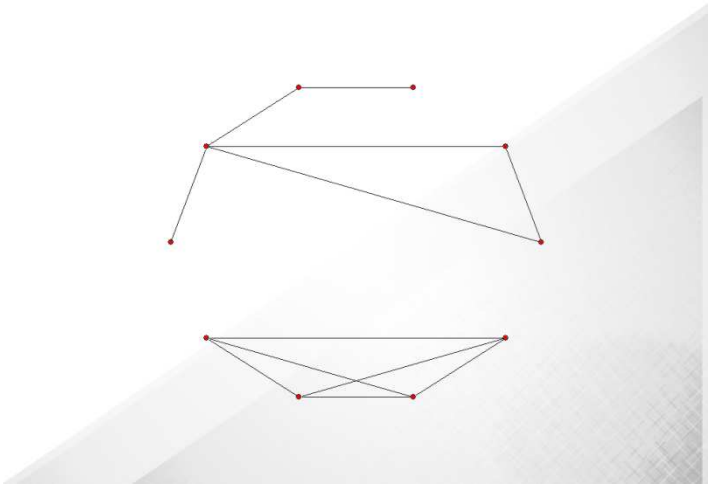
2.3 Representando e Comparando o Léxico Mental

Na presente pesquisa, optou-se por investigar o léxico-mental de grupos de diferentes faixas etárias. Esse tipo de comparação é importante, uma vez que o desenvolvimento lingüístico é caracterizado por fases específicas. Por exemplo, aos 5 meses a criança já conhece uma grande quantidade de conceitos, mas somente começará a empregá-los por volta dos 12 meses (TOMASELLO, 2003). Entre 2 e 3 anos, as crianças já são capazes de expressar-se verbalmente, mas ainda encontram-se em amplo aprendizado gramatical. Quando são expostas a sentenças com formas gramaticais incorretas, elas tendem a imitá-las (num processo de experimentação), enquanto crianças de 4 anos tendem a corrigi-las (GOLDBERG, 1999). Entre 3 e 4 anos, é possível perceber o fenômeno da generalização excessiva (*overgeneralization*), quando as crianças aplicam inadeqüadamente estruturas aprendidas em verbos conhecidos, buscando aumentar seu conhecimento sem depender do *input* (GOLDBERG, 1999). Aos 5 anos, considera-se que a criança possua um vocabulário considerado complexo, mas somente aos 10 ele estaria efetivamente consolidado (JACKENDOFF, 1994). As fases apresentadas podem ser agrupadas em três momentos específicos: aprendizado e teste (até o 3 anos), maturação (por volta dos 5 anos) e domínio completo (adultos).

Neste trabalho, foram utilizados dados de grupos que representam as três fases mencionadas (como será visto no capítulo 3). Para realizar as comparações, o léxico-mental de cada grupo foi representado por um grafo. A opção por esta representação deve-se à facilidade de expressar relações semânticas entre os componentes, como a sinonímia (STEYVERS & TENENBAUM, 2005) e homonímia (GORMAN & CURRAN, 2007). Para referências de outros trabalhos utilizando grafos, consultar a seção 2.1.1.

Objetivamente, um grafo G consiste em um conjunto de vértices (ou ‘nodos’) $V = \{1, ..., v\}$, com arcos $E = \{1,..., e\}$ ligando-os. Dados dois nodos quaisquer i e $j \in V$, um conjunto de arcos ligando i a j é chamado ‘caminho’. O ‘tamanho’ do caminho corresponde à quantidade de arcos que o compõe. O menor caminho entre i e j é chamado ‘caminho mínimo’ (ou ‘geodésico’). Todos os nodos que possuem arcos incidentes de i são denominados ‘vizinhos de i’. O conjunto dos vizinhos de i, ou a ‘vizinhança de i’ é denotada por N_i, enquanto o tamanho dessa vizinhança, também conhecido como ‘grau’ do nodo, é denotado por k_i.

Os grafos serão avaliados por meio de métricas da teoria dos grafos comumente adotadas na literatura (explicadas a seguir): *S*, *C/s*[8], *L*, *diam*, <*k*> e *D*. Dentre os autores que já as utilizaram, destacamos Motter et al. (2002), utilizando *C*, *L* e <*k*>; Sigman e Cecchi (2002), utilizando *C* e *L*; Ferrer-i-Cancho et al. (2004), utilizando *C*, *L* e <*k*>; Steyvers e Tenenbaum (2005), utilizando *C*, *L*, *diam* e <*k*>; Masucci e Rodgers (2006), utilizando *C*, e <*k*>; Gaume et al. (2006), utilizando *C*, *L*, <*k*> e *D*; Gorman e Curran (2007) , utilizando *C*, *L*, *diam* e <*k*>; De Deyne e Storms (2008), utilizando *C*, *L*, *diam*, <*k*> e *D*. Um exemplo de grafo com as respectivas métricas associadas pode ser visto na figura 2.1:



Dados: V = 10, E = 12
Métricas: *S* = 2 *C/s* = 0,41 *L* = 1,57 *diam* = 3 <*k*> = 2,4 *D* = 0,27

Figura 2.1: Um grafo e suas métricas.

Além de métricas de teoria dos grafos, a presente pesquisa inova ao utilizar também medidas da teoria dos conjuntos. A motivação para a utilização de dois tipos diferentes de métricas (de teoria dos grafos e de teoria dos conjuntos) deu-se pela necessidade de avaliar de múltiplas formas os mesmos dados. A análise topológica da teoria dos grafos justifica-se por permitir a verificação de similaridades estruturais, avaliando puramente a organização da informação, desconsiderando o conteúdo. Dessa forma, mesmo que utilizando termos distintos, é possível verificar se o modo como os grupos de indivíduos expressam informações é similar (figura 2.2). Afinal, ainda que possivelmente utilizando palavras diferentes, é interessante poder verificar se o modo como os grupos estruturam as informações é similar.

[8] A métrica *C/s* é uma variação da métrica *C*, como será explicado a seguir.

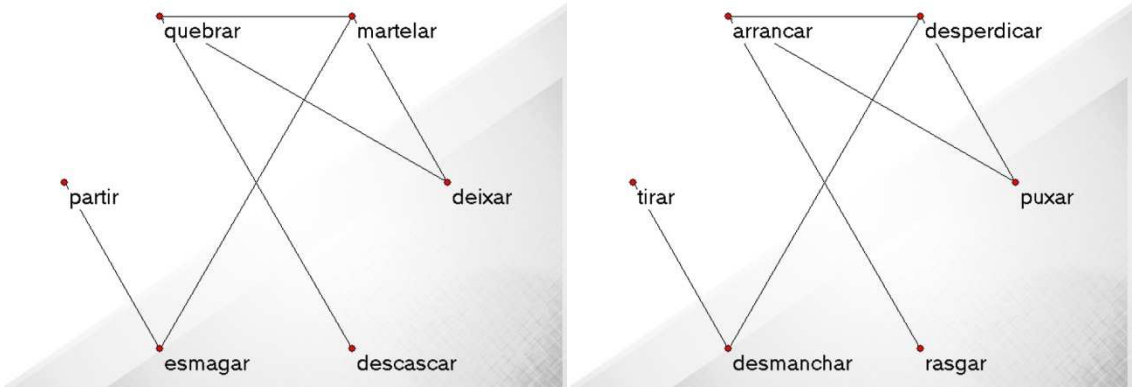


Figura 2.2: Dois grafos com mesma estrutura e conteúdos diferentes.

Por outro lado, a teoria dos conjuntos permite-nos focar no conteúdo puro, desconsiderando a estrutura. Mesmo que a organização seja distinta, é possível verificar se o as representações das informações (as palavras) coincidem nos diferentes grupos (figura 2.3).

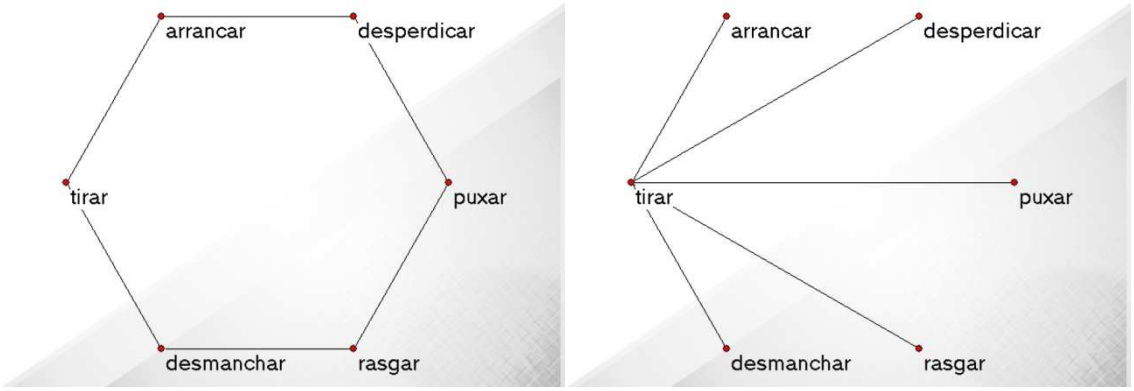


Figura 2.3: Dois grafos com mesmo conteúdo e estruturas diferentes.

2.3.1 Teoria dos Grafos: Verificando a Similaridade de Estrutura

As principais medidas de teoria dos grafos utilizadas nesta pesquisa são:

- Quantidade de subgrafos (*S*);
- Coeficiente de clusterização (*C/s*);
- Tamanho do caminho mínimo médio (*L*);
- Diâmetro (*diam*);
- Conectividade média (<*k*>);
- Densidade (*D*).

A métrica "quantidade de subgrafos" (*S*) refere-se à quantidade de grafos desconexos existentes em um determinado grafo. Ela permite a verificação da uniformidade do grafo, refletindo a estabilidade de sua estrutura. Ainda que não seja uma métrica comum, foi aqui incluída, uma vez que será preciso avaliar também grafos desconexos.

O coeficiente de clusterização (*C*) representa o quão clusterizado é o grafo, ou seja, mede a existência de agrupamentos onde os vértices participantes possuem um alto índice de conectividade entre si. Com esta métrica procuramos medir a estrutura do grafo como um todo, em especial a distribuição relativa dos arcos. Foi utilizada a medida local apresentada por Watts e Strogatz (1998), aplicada sobre todo o grafo. A medida local é a proporção de arcos existentes entre os vizinhos de i, dividido pela quantidade máxima de arcos possíveis entre eles. Sendo, j e $k \in N_i$ e $e_{jk} \in E$, o coeficiente de clusterização local de i pode ser obtido por:

$$C_i = \frac{2|\{e_{jk}\}|}{(k_i(k_i-1))} \quad .$$

Os valores de C_i variam entre 0 (quando seus vizinhos não possuem arcos entre si) e 1 (quando todos os vizinhos estão completamente interconectados). Obtém-se o valor para o grafo como um todo por meio da média dos coeficientes locais:

$$C = \frac{\sum_{i=1}^{n} C_i}{n} \quad .$$

Ainda que *C* seja uma medida amplamente conhecida e adequada para uma avaliação estrutural, ela não é indicada para grafos desconectados. Uma grafo composto unicamente por vértices isolados (sem arestas), por exemplo, resultaria em $C = 1$, uma vez que, não existem vizinhos que não estejam conectados. O mesmo acontece com grafos compostos apenas por pares de nodos conectados, ou triplas, etc.

Dessa forma, para levar em conta a desestruturação do grafo tanto por conta da diminuição na clusterização, quanto por conta de particionamento, *C* foi modificada. A métrica *C/s* resultante consiste na métrica *C* dividida pela quantidade de subgrafos desconexos existentes (*S*):

$$C/s = \frac{C}{S} \quad .$$

Dessa forma, um grafo composto por dois agrupamentos separados de nodos totalmente conectados entre si, tem $C/s = ½ = 0{,}5$.

A conectividade média (<*k*>) corresponde à média das conectividades dos vértices; à proporção entre vértices e arcos. Novamente, o objetivo é verificar a estrutura do grafo: valores maiores tendem a indicar agrupamento de vértices (mais próximo de um grafo completo), enquanto valores menores tendem a indicar espaçamento. Pode ser calculada por:

$$<k> = \frac{\sum_{i=1}^{n} k_i}{n} = \frac{|E|}{|V|} \quad .$$

A densidade (*D*) corresponde à quantidade de arcos existentes em relação ao total possível para a quantidade de vértices. É uma medida similar à "<*k*>", uma vez que envolve arcos e vértices, mas é mais precisa. Em vez de expressar a proporção direta entre esses dois fatores, mede a quantidade de arcos existentes em relação ao máximo possível para a quantidade de vértices do grafo. Esta é uma variação não-linear, uma vez que o acréscimo de um vértice ($|V| = n + 1$) leva a um aumento significativo na quantidade máxima de arestas, proporcional à nova quantidade de vértices ($Max(|E|)' = Max(|E|)(n+1)$). Pode ser calculada por:

$$D = \frac{2|E|}{|V|(|V|-1)} \quad .$$

O caminho mínimo médio (*L*) corresponde à média das distâncias mínimas de todos os vértices para todos os outros vértices, expressando, portanto, o tamanho do grafo como um todo. Com esta métrica, objetiva-se verificar a estrutura do grafo, em especial a distribuição relativa dos vértices. Seja dg_{ij} a distância mínima entre quaisquer vértices *i* e *j*, o caminho mínimo médio do grafo pode ser calculado por:

$$L = \frac{2\sum_{i=1}^{n-1}\sum_{j=2}^{n} dg_{ij}}{(n(n-1))}, \text{ tal que } i > j.$$

Uma vez que a métrica terá de ser calculada também sobre grafos desconexos, os caminhos entre nodos não alcançáveis foram desconsiderados.

O diâmetro (*diam*) mensura os grafos de forma similar a *L*, mas focando em um aspecto mais simples. Enquanto *L* expressa o tamanho do grafo, o diâmetro expressa sua maior extensão, ou seja, é o maior caminho mínimo:

$$diam = max(\{dg_{ij}\}), \text{ para quaisquer nodos } i \text{ e } j.$$

A escolha de <*k*> e *D* objetiva mensurar indiretamente o compartilhamento semântico do grafo. Uma vez que cada ligação indica uma intersecção de significado entre os verbos envolvidos (como será visto no capítulo 3), verificar a quantidade de ligações permite-nos medir o quão semanticamente consistente é o conjunto de verbos. Uma quantidade maior de ligações tende a indicar que os verbos são semanticamente próximos: foram empregados nas mesmas situações múltiplas vezes, portanto suas definições coincidem. *L* e *diam* permitem verificar a uniformidade do grafo: quanto maiores os valores, mais esparso ele tende a ser. Um grafo esparso indica a formação de centros de significado isolados uns dos outros, com poucos nodos *hub* interligando os grupos. A baixa quantidade de *hubs*, por sua vez, tende a indicar um vocabulário mais específico, com palavras menos gerais (que acabam podendo ser utilizadas em poucas situações diferentes). Em outras palavras, *L* e *diam* permitem verificar também a especificidade do vocabulário. *S* permite verificar os casos extremos de esparsidade, quando um dos grupos torna-se tão isolado que deixa de estar ligado com o restante do grafo, gerando uma divisão. Nesse caso, *L* e *diam* serão baixos (a distância entre os vértices alcançáveis diminui), mas *S* aumenta. Por fim, *C/s* realiza uma medição mais completa, verificando a uniformidade tanto de grafos conexos (por meio da medição da clusterização), quanto de grafos desconexos (por ter o valor dividido por *S*).

2.3.2 Teoria dos Conjuntos: Verificando a Similaridade de Conteúdo

A teoria dos conjuntos permite-nos verificar a similaridade do conteúdo entre dois agrupamentos de dados. São métricas comparativas, ou seja, são necessários dois conjuntos para aplicá-las. Para utilizar estas medições, os grafos comparados foram considerados como simples listas de verbos, ignorando-se as ligações entre os vértices.

As métricas utilizadas baseiam-se no coeficiente de Jaccard (JACCARD, 1901). Dados dois conjuntos *A* e *B*, o coeficiente de Jaccard *J* pode ser calculado da seguinte forma:

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

onde "*x*" é o número de elementos presentes em ambos *A* e *B*, ou seja,

$$x = A \cap B ,$$

"*y*" é o número de elementos presentes apenas em *A*, ou seja,

$$y = A - B ,$$

e "*z*" é o número de elementos presentes apenas em *B*, ou seja,

$$y = B - A \ .$$

Nas simulações computacionais realizadas (explicadas no capítulo 3), os conjuntos correspondem às listas de verbos de cada grafo, sendo que *A* é assumido como o conjunto do grafo de referência, e *B* como o conjunto do grafo sendo modificado.

Nas simulações, além do próprio coeficiente de Jaccard, serão apresentadas a quantidade de verbos excluídos comuns aos dois grupos (a componente "*x*" da fórmula) e a quantidade de verbos excluídos diferentes, presentes apenas em *B* (a componente "*z*" da fórmula). Ainda que o coeficiente de Jaccard seja uma medida conhecida, a análise de suas componentes facilita a análise dos resultados, como será verificado mais adiante.

3 O MODELO

A presente pesquisa investiga a evolução do léxico-mental verbal por meio de simulações sobre conjuntos de dados, que representam o léxico-mental de grupos de indivíduos. Cada conjunto é expresso por um grafo, cujos vértices representam verbos dados como resposta em um experimento psicolingüístico. As simulações realizadas na presente pesquisa consistem em modificações sobre esses conjuntos de dados, visando tornar um mais similar ao outro. Essa aproximação é feita por meio da eliminação progressiva dos vértices, ordenadamente, de acordo com os fatores lingüísticos testados (polissemia, freqüência e complexidade sintática). A similaridade é verificada por meio de métricas da teoria dos grafos e da teoria dos conjuntos. Neste capítulo, serão apresentados os dados utilizados e uma descrição do procedimento de simulação.

3.1 Dados

Os grupos utilizados na presente pesquisa consistem de 55 crianças e 55 adultos jovens[9]. Para que o estudo englobasse a evolução em crianças, os dados delas são longitudinais, com uma segunda coleta depois de dois anos (TONIETTO et al., 2008). Assim, a idade dos participantes da primeira coleta (G1) varia entre 2:0 e 3:11 (média de 3 anos e 1 mês) e entre 4:1 e 6:6 (média de 5 anos e 5 meses) na segunda (G2). O grupo dos adultos (G3) é composto por indivíduos entre 17 e 34 anos (média de 21 anos e 8 meses), não relacionados às crianças. Estes dados nos permitiram comparar a evolução no léxico tanto durante a idade de aquisição (entre G1 e G2) quanto em relação aos adultos (entre G2 e G3). Os grupos coincidem com os três momentos importantes do desenvolvimento lingüístico mencionados na seção 2.3. Os participantes[10] são falantes do Português Brasileiro, o que torna a pesquisa bastante particular, uma vez que são poucos os estudos lingüístico-computacionais neste idioma.

Os conjuntos de dados foram obtidos por meio de uma tarefa psicolingüística de nomeação. Cada participante assistiu a dezessete vídeos (cada um com a duração aproximada de 1 minuto) de ações de divisão ou destruição. Cada vídeo iniciava com uma mulher pegando um objeto (dentre vários que existiam sobre uma mesa) e, a seguir, desempenhando uma ação (rasgando um jornal ou descascando uma laranja, por exemplo). Por fim, o objeto era filmado por alguns segundos. Depois da exibição de

[9] Originalmente, foram coletadas informações de 75 adultos. Entretanto, para evitar que a diferença no número de integrantes interferisse nos resultados, 55 deles foram escolhidos aleatoriamente para participar da presente pesquisa.

[10] As crianças (32 meninos e 23 meninas) são oriundas de instituições educacionais particulares da Região Metropolitana de Porto Alegre. Não foram incluídas crianças bilíngües, nem com alguma dificuldade física ou cognitiva que pudesse influenciar na comparação com outras na mesma faixa-etária e em mesmas condições socioeconômicas (TONIETTO, 2009).

cada vídeo, foi perguntado aos participantes “O que a mulher fez?” (como descrito em TONIETTO, 2009). O domínio ‘divisão e destruição’ foi escolhido devido a sua importância cognitiva: ele corresponde a uma das quatro zonas conceituais principais, agrupando uma grande quantidade de verbos (como descrito em TONIETTO, 2009). As outras zonas são ‘evasão’, ‘excitação’ e ‘união’.

O resultado da tarefa de nomeação foi uma tabela contendo uma resposta de cada indivíduo para cada vídeo, num total de 2805 respostas (935 por grupo). Sobre estas respostas, foi executado um procedimento de ‘limpeza’, de modo a manter apenas as respostas válidas e formatadas. Assim, respostas contendo complementos, mas cujo verbo era evidente, foram apenas transformadas: “rasgar um jornal” em “rasgar”, “descascar com a mão” em “descascar”, “tirar as peças” em “tirar”, etc. Respostas inválidas como “não sei”, “hum” e demonstrações não verbais (“fazer assim” e mostrar com as mãos o resultado) foram eliminadas. A seguir, um segundo procedimento de ‘limpeza’ foi executado, de modo a eliminar as respostas mencionadas apenas uma vez (e que dificilmente expressariam o vocabulário do grupo todo). O impacto dessa manipulação dos dados sobre a quantidade de respostas e o tamanho do gráfico pode ser visto nas tabelas 3.1 e 3.2 respectivamente. A média de respostas por vídeo e a média de respostas diferentes por vídeo (ambas coletadas após a segunda ‘limpeza’) podem ser vistas na tabela 3.3. A média de respostas por vídeo consiste na divisão simples do total de respostas do grupo pela quantidade de vídeos (17), enquanto a média de respostas diferentes é resultado da análise dos subgrafos de cada vídeo em separado (ver seção 3.1.1). Uma vez que muitos verbos foram mencionados em múltiplos vídeos, este último resultado é muito maior do que a simples divisão da quantidade de verbos diferentes mencionados pelo grupo divididos pela quantidade de vídeos.

Tabela 3.1: Impacto dos procedimentos de limpeza nas respostas dos grupos.

	G1	G2	G3	Total
Quantidade de respostas por grupo	935	935	935	2805
Quantidade de respostas após a primeira limpeza	794	929	932	2655
Quantidade de resposta após a segunda limpeza	785	911	917	2613

Tabela 3.2: Impacto dos procedimentos de limpeza nos grafos dos grupos.

	G1	G2	G3
Tamanho do grafo após a primeira limpeza	31	43	46
Tamanho do grafo após a segunda limpeza	22	25	31
Redução do tamanho entre limpezas	29,03%	41,86%	32,61%

Tabela 3.3: Média de respostas depois da segunda limpeza.

	G1	G2	G3
Média de respostas por vídeo	46,18	53,59	53,94
Média de respostas diferentes por vídeo	6,76	5,53	4

3.1.1 Grafos

A partir das respostas individuais devidamente limpas, foi construído um grafo do vocabulário do grupo, levando em consideração os vídeos em que cada verbo foi mencionado. Foram criados, primeiramente, subgrafos, um para cada vídeo, onde todos os verbos utilizados para descrever a cena vista foram ligados uns aos outros (um clique). Estes subgrafos foram então unidos (*merging*) por meio dos verbos mencionados em múltiplos vídeos, denominados *hubs*, gerando um grafo global, que representa o vocabulário de todo o grupo em relação a todos os vídeos (figura 3.1).

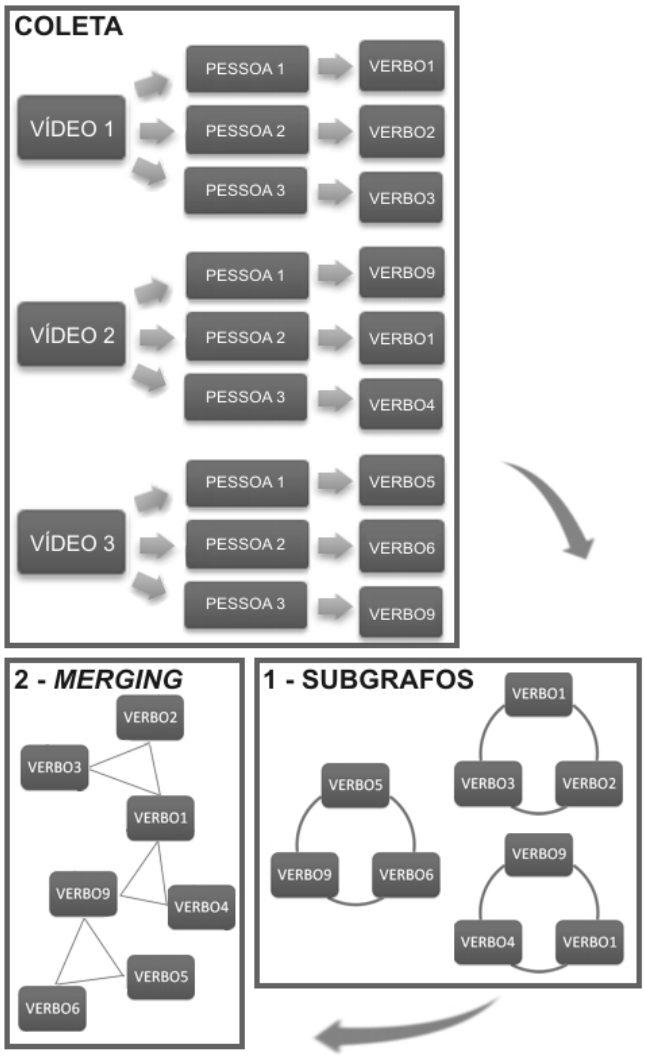


Figura 3.1: Processo de construção dos grafos de cada grupo. Inicialmente foi feita a coleta dos verbos, por meio da tarefa psicolingüística, e as 'limpezas'. Os grafos foram construídos com base nesses verbos, inicialmente um grafo por vídeo. A seguir, foi construído o grafo do vocabulário do grupo como um todo, fazendo-se um *merging* dos verbos que aparecem em mais de um subgrafo.

Ainda que seja difícil verificar com precisão, por conta da pequena quantidade de nodos, os grafos resultantes assemelham-se a redes do tipo *small-world*. Foi demonstrado que diversos tipos de redes complexas apresentam essa estruturação, como a WWW, as redes de relacionamento pessoal (tendo as pessoas como vértices e as amizades como arestas) e a rede elétrica (WATTS & STROGAZ, 1998; ALBERT & BARABÁSI, 2002; STEYVERS & TENENBAUM, 2005; GAUME et al., 2006). Grafos léxicos de diversos tipos também apresentam essa caracterização (GAUME et al., 2006).

Grafos *small-world* apresentam três características principais: grande coeficiente de clusterização (*C*), baixo caminho mínimo médio (*L*) e uma distribuição *power-law*[11] dos graus dos vértices (<*k*>). A presença de um grande *C* é evidente ao compararmos os grafos obtidos com outros gerados aleatoriamente, com a mesma densidade (tabela 3.4). Entretanto, o *L* dos grafos aleatórios demonstrou ser um pouco menor, o que contraria os requisitos de *small-world*. Quanto à distribuição de <*k*>, ela não pode ser medida com precisão, uma vez que demandaria algumas centenas (preferencialmente milhares) de vértices para que pudesse ser examinada adeqüadamente.

Tabela 3.4: Comparação dos grafos obtidos com grafos aleatórios de mesma densidade.

Métrica	G1	G1-aleatório	G2	G2-aleatório	G3	G3-aleatório
Coeficiente de clusterização	0,84	0,59	0,78	0,42	0,78	0,3
Caminho mínimo médio	1,46	1,45	1,60	1,59	1,98	1,74

A caracterização de um grafo *small-world* dá-se pela grande localidade das ligações e pela existência de vértices *hub* ligando porções 'distantes' do grafo (STEYVERS & TENENBAUM, 2005). Ainda que o método de construção apresentado favoreça tanto a localidade quanto a existência de *hubs*, esta quantidade não foi suficiente para que houvesse um efetivo decréscimo de *L*. Neste sentido, é preciso lembrar que a proporção de *hubs* depende, na presente pesquisa, da similaridade entre os vídeos apresentados: quanto mais similares, maior a tendência de que a mesma palavra seja utilizada para descrever múltiplos vídeos. Mesmo que apresentando um domínio comum (divisão ou destruição), este é um fator importante a ser considerado.

Os grafos obtidos podem ser vistos na figura 3.2. Em G1, existe uma grande quantidade de vértices com muitas ligações[12], o que resulta num espalhamento pequeno de vértices. Em G2, a quantidade de vértices com poucas ligações é maior, uma tendência que aumenta em G3. A definição do grafo de cada grupo (incluindo os vértices presentes e as ligações entre eles) encontra-se no anexo B.

[11] Uma distribuição *power law* pode ser facilmente identificada por uma linha reta na plotagem de coordenadas log-log.

[12] A grande quantidade de ligações no todo e a grande concentração de ligações em alguns nodos (grande desvio padrão) podem ser comprovadas na tabela 4.1.

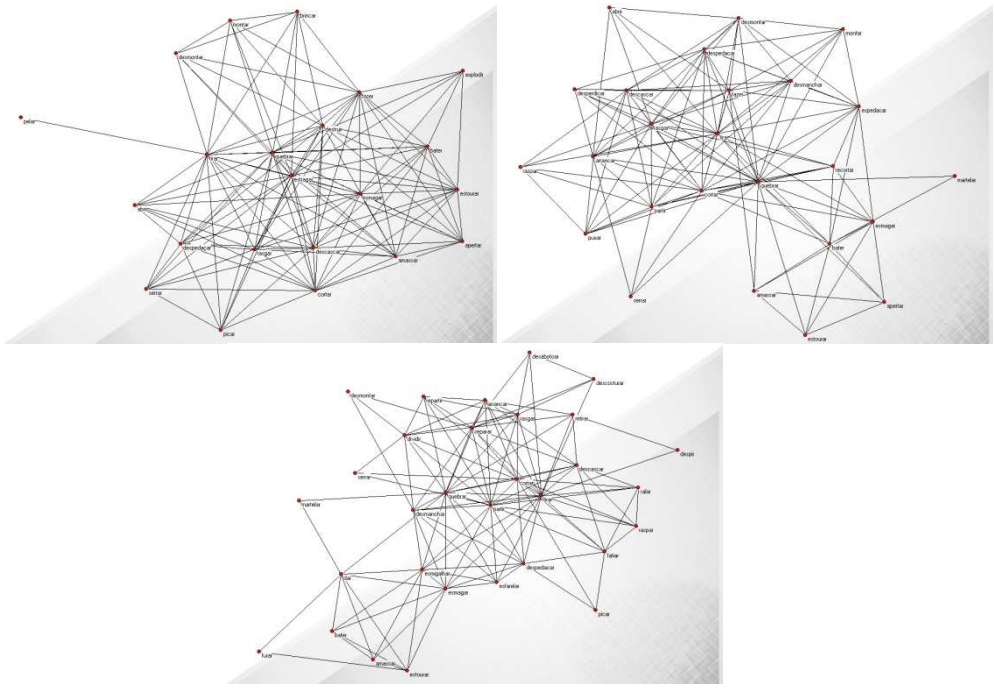


Figura 3.2: Grafos dos grupos G1, G2 e G3 (respectivamente). Foram plotados com a ferramenta Pajek (BATAGELJ & MRVAR, 2009).

A motivação para relacionar entre si todos os verbos mencionados para descrever um vídeo foi a hipótese de compartilhamento de significado. Cada filme prove um contexto semântico com diversas possibilidades de escolha lexical, onde cada palavra diferente descreve alguma característica do contexto. Assim, assume-se que palavras utilizadas para descrever o mesmo contexto estejam relacionadas semanticamente. Mais do que uma suposição, a existência de relações conceituais motivadas por associações semânticas está em conformidade com Nelson et al. (1998), que demonstraram que relações semânticas implícitas (entre diferentes palavras) influenciam nos processos de reconhecimento e recordação. No contexto da presente pesquisa, as ligações entre os vértices são as relações semânticas implícitas.

A metodologia de construção dos subgrafos (relacionando todos os verbos que descrevem o mesmo vídeo) foi inspirada no sistema conceitual humano, mais precisamente na definição de ‘semelhanças de família’ de Wittgenstein (1953) e no princípio da ‘figura-fundo’, da teoria Gestalt (mais detalhes em EVANS & GREEN, 2006; SCHMID & UNGERER, 2006). Segundo Wittgenstein, a definição de um conceito não é composta simplesmente por um conjunto necessário e suficiente de características. O conceito ‘jogo’, por exemplo, pode englobar características bastante diversas, como a existência ou não de competição, com múltiplas pessoas ou apenas uma (por exemplo, ‘futebol’ ou o jogo de cartas ‘paciência’, respectivamente, em ambos os casos), demandar sorte (como nos jogos de dado) ou apenas habilidade (como no ‘xadrez’ ou ‘damas’). Ainda que a quantidade de características potenciais associadas a ‘jogo’ seja grande, parece não ser possível encontrar uma que esteja realmente presente em todos os tipos de jogos. O conceito ‘jogo’ é, portanto, definido por ‘semelhanças de família’, onde diversos grupos de subconceitos (neste caso os tipos de jogos) compartilham conjuntos de características. Os subconceitos acabam agrupados (como tipos de jogos) não por compartilharem um conjunto de características, mas por relacionarem-se fortemente uns aos outros.

Quando nos referimos ao conceito ‘jogo’, nas mais diversas situações, estamos, implicitamente, valorizando algumas das características possíveis. Dentro do contexto ‘Olimpíada’, por exemplo, os ‘jogos’ referem-se às competições esportivas, enquanto no contexto ‘ludoteca’, os ‘jogos’ tenderão a referir-se a brinquedos. A valorização de alguns aspectos de um todo é chamada de princípio da ‘figura-fundo’. Componente da teoria Gestalt, esse princípio define que, em determinados momentos, segundo alguns critérios, algumas partes são simplesmente consideradas cognitivamente mais importantes do que outras.

Analogamente, durante o processo de construção dos subgrafos, o vídeo pode ser visto como o ‘contexto’. Esse contexto gera uma valorização de algumas características das palavras candidatas como resposta, ou seja, trata-se da aplicação do princípio da ‘figura-fundo’ na busca pela melhor descrição. A palavra escolhida será aquela cujas características valorizadas melhor descreverem o estímulo na opinião do indivíduo sendo avaliado. Uma vez que múltiplas palavras foram utilizadas para descrever o mesmo vídeo, assumimos que as características valorizadas em cada resposta são as mesmas (ou, ao menos, bastante próximas). Em outras palavras, da mesma forma que alguns dos múltiplos significados de ‘jogo’ compartilham porções de significado, assumimos que as palavras utilizadas na descrição do vídeo também o fazem, sendo, portanto, relacionadas.

3.2 Os Fatores Lingüísticos

Os fatores lingüísticos escolhidos serão avaliados quanto a seu impacto no processo evolutivo. Segundo o que foi exposto no capítulo 2, podemos consolidar os resultados esperados nas seguintes hipóteses.

- Crianças tendem a aprender primeiro os verbos mais freqüentes.
- Crianças tendem a aprender primeiro os verbos mais polissêmicos.
- Crianças tendem a aprender primeiro os verbos com menor complexidade sintática.
- A organização da informação evolui de acordo com estes fatores.

Para verificar as hipóteses, cada verbo mencionado na tarefa de nomeação foi anotado com escores representativos dos fatores lingüísticos, que serão utilizados no processo de simulação. Os escores de polissemia foram coletados a partir de duas fontes distintas. A primeira foi o ‘WordNetBR’ (DIAS-DA-SILVA et al, 2000; DIAS-DA-SILVA & MORAES; 2003; MAZIERO, 2008), a versão em Português brasileiro do conhecido tesauro eletrônico ‘WordNet’. Polissemia, nesse contexto, foi extraída contando-se a quantidade de *synsets* (*synonym sets*, ou, conjuntos de sinônimos) diferentes na qual um verbo participava[13]. A ferramenta CLAN (MACWHINNEY, 2000) foi utilizada para auxiliar na automação dessa tarefa. A segunda fonte de consulta foi a versão eletrônica do dicionário Houaiss (HOUAISS, 2007). Nesse caso, a quantidade de significados de cada verbo foi contabilizada manualmente.

A coleta dos escores da freqüência de observação dos verbos também foi feita a partir de duas fontes de dados. A primeira delas foi o buscador ‘Yahoo!’ (por meio de

[13] O verbo ‘serrar’ não constava no WordNetBR no momento da coleta, portanto foi atribuído a ele o valor ‘0’.

um programa utilizando a API disponibilizada[14]), onde cada verbo foi buscado no infinitivo, em páginas localizadas no Brasil e em português. Essa busca permitiu verificar a utilização global e geral de cada verbo. A segunda fonte foi o corpus 'Florianópolis' (SCLIAR-CABRAL, 1993; MACWHINNEY, 2000), composto por 5530 frases, coletadas em três sessões, totalizando 15 horas de conversas entre adultos e uma criança. Os verbos foram extraídos com o auxílio da ferramenta CLAN (MACWHINNEY, 2000) e depois filtrados manualmente[15] para a eliminação de ruído (emprego na forma substantivada ou palavras similares coletadas por engano). Tanto frases das crianças quanto dos adultos foram utilizadas na contabilização da freqüência.

Dos 44 verbos mencionados pelos três grupos, 25 não foram encontrados no corpus 'Florianópolis', recebendo, portanto, o valor '0' como freqüência associada. A ausência de verbos era esperada, uma vez que estamos analisando um contexto bastante particular (divisão e destruição) e o corpus apresenta uma interação geral, não atrelada a nenhum tema específico. De qualquer forma, uma vez que o objetivo da coleta era justamente verificar a freqüência de utilização das palavras, a ausência sugere uma freqüência baixa.

Os escores da complexidade sintática foram extraídos também a partir do corpus 'Florianópolis', com o auxílio da ferramenta CLAN (MACWHINNEY, 2000). Depois da coleta, cada frase foi avaliada (pelo autor da presente pesquisa e, posteriormente, por um especialista em lingüística) segundo a quantidade e o tipo de objetos relacionados ao verbo. O valor '0' foi atribuído quando o verbo não possuía nenhum objeto (verbo intransitivo), '1' quando possuía um objeto direto (verbo transitivo direto), '2' quando possuía um objeto indireto (verbo transitivo indireto) e '3' quando possuía, ao mesmo tempo, objeto direto e objeto indireto (verbo bitransitivo)[16]. Uma vez que a linguagem analisada é coloquial (portanto sujeita a erros gramaticais), foi preciso criar regras específicas para casos excepcionais, de forma similar a Laakso e Smith (2007):

1. Frases compostas unicamente pelo verbo (como "Tira.") ou outros complementos que não o objeto (como "Tira aqui." ou "Eu qué tirá."), mesmo que eventualmente referindo-se a um objeto anteriormente mencionado, foram consideradas como contendo um verbo intransitivo.
2. Frases na voz passiva, i.e., que descrevem uma ação (por exemplo, "Ele dormiu até as cinco horas, né."), foram consideradas como contendo um verbo intransitivo.
3. Frases onde não foi possível compreender se uma palavra poderia ou não ser considerada um objeto (como "Deixa ele abrir xxx." ou "Deixa u@fp caixa.") foram ignoradas.

Sabendo-se que as frases foram extraídas do corpus 'Florianópolis', foi preciso classificar os 25 verbos não encontrados de alguma forma. Desse modo, o valor '20' foi arbitrariamente atribuído a cada um deles, simbolizando uma dificuldade extrema na utilização. A motivação para esta decisão foi manter a conformidade com a hipótese

[14] http://developer.yahoo.com/search/.

[15] Uma vez que as variações das formas verbais no português são muitas, foi preciso filtrar a partir da maior porção comum a todas as formas e depois manualmente excluir os falso-positivos encontrados: "cortar" procurou por 'cort*', 'martelar' procurou por 'martel*', 'dar' procurou por 'd*', etc.

[16] A seção 2.2.3 apresenta as definições dos termos entre parênteses.

assumida: quanto maior a complexidade sintática, mais difícil será para a criança dominar e utilizar determinada palavra. Assim, assumimos que os verbos não mencionados estão associados a uma dificuldade de utilização, e que, portanto, teriam uma complexidade maior.

Para definir o valor efetivamente associado a cada verbo, foram analisadas as frases mencionadas apenas pelos adultos: a regência verbal (quantidade de complementos) mais utilizada foi a que teve seu valor associado ao verbo. A opção por ignorar as frases da criança deve-se ao fato de que elas estariam excessivamente sujeitas a erros gramaticais. Além disso, a configuração de objetos utilizada pela criança não é necessariamente aquela a que ela está exposta com maior freqüência.

Ainda que as respostas da criança não tenham sido utilizadas na simulação, elas foram coletadas e analisadas, de forma que pudessem ser comparadas com os resultados obtidos pelos adultos. A primeira diferença perceptível é uma maior quantidade de verbos com o valor ‘20’ associado, o que era esperado, uma vez que uma única criança contribuiu com muito menos frases do que múltiplos adultos (portanto mencionou muito menos verbos). A segunda diferença é uma clara tendência em utilizar a forma verbal intransitiva, comumente indicando os objetos a que o verbo se refere de forma física (apontando, por exemplo), enquanto os adultos verbalizam mais completamente suas idéias. As diferenças podem ser verificadas na tabela C.3 do Anexo C.

Além dos fatores lingüísticos simples, foram criadas duas combinações, com o objetivo de verificar se produziriam resultados melhores do que suas componentes originais. Uma vez que se espera que múltiplos fatores influenciem no processo de evolução, é possível que a combinação venha a produzir um resultado otimizado. Cada combinação consistiu na união da freqüência extraída do ‘Yahoo!’ com uma das duas polissemias. Os dois fatores extraídos do corpus ‘Florianópolis’ foram ignorados por conta do grande número de palavras não mencionadas.

O processo de combinação desconsiderou o valor bruto associado aos verbos por cada parâmetro, uma vez que os valores da freqüência tendem a ser muito maiores do que os da polissemia. Assim, optou-se por, inicialmente, criar listas ordenadas pelos valores de cada fator lingüístico e, a seguir, utilizar a posição da palavra nas listas no procedimento de combinação. Em outras palavras, a cada verbo, foi associado ao valor de sua posição: ‘1’ para o primeiro, ‘2’ para o segundo, etc. Esta posição foi determinada apenas pelos verbos mencionados dentro do grupo (não foi criado um *ranking* geral englobando verbos dos três grupos). Quanto aos empates, todos os verbos foram considerados como estando na mesma posição, recebendo, portanto, o mesmo número. O valor do verbo na combinação é resultado da média dos valores ordinais em ambos os fatores lingüísticos. Os valores associados a cada verbo pelas combinações, bem como pelos fatores lingüísticos encontra-se no apêndice C.

O procedimento de combinação é mostrado na figura 3.3. Inicialmente, os verbos de um grupo foram arranjados em ordem crescente de polissemia e de freqüência (separadamente). Cada verbo foi associado a uma posição nessas ordenações. O valor considerado para a combinação é a média das posições nas duas ordenações iniciais, e o resultado é uma lista ordenada por esta média. Esta será a ordem de eliminação na fase de simulação computacional.

Verbo	Polissemia	Posição
VERBO2	5	1
VERBO1	9	2
VERBO5	9	2
VERBO4	17	3
VERBO3	32	4

Verbo	Freqüência	Posição
VERBO1	1000	1
VERBO2	12000	2
VERBO4	13500	3
VERBO3	120000	4
VERBO5	2340000	5

Verbo	Combinação	Posição
VERBO1	(2+1)/2 = 1,5	1
VERBO2	(1+2)/2 = 1,5	1
VERBO4	(3+3)/2 = 3	2
VERBO5	(2+5)/2 = 3,5	3
VERBO3	(4+4)/2 = 4	4

Figura 3.3: Procedimento de combinação.

Visando facilitar as referências aos diferentes valores medidos para os fatores lingüísticos e combinações, foram criados nomes representativos para eles:

- PolWord: valores de polissemia extraídos do WordNetBR.
- PolHou: valores de polissemia extraídos do dicionário eletrônico Houaiss.
- FreqYahoo: valores de freqüência de observação extraídos da API do buscador ‘Yahoo!’.
- FreqFlorian: valores de freqüência de observação extraídos do corpus ‘Florianópolis’.
- CompSint: valores de complexidade sintática
- CombYahooWord: valores da combinação entre FreqYahoo e PolWord.
- CombYahooHou: valores da combinação entre FreqYahoo e PolHou.

3.3 A Simulação

As simulações consistem em modificações de um grafo de modo a testar a influência dos fatores lingüísticos no processo evolutivo da linguagem. O objetivo central é modificar o grafo de um dos grupos etários, de modo a torná-lo mais parecido com o de outro grupo (simulando a evolução). O processo foi baseado em estratégias de modificação topológica como o *network growth*, ou crescimento da rede (ALBERT & BARABÁSI, 2002), que têm sido utilizadas para testar a influência de fatores específicos na estruturação da linguagem (e.g. STEYVERS & TENENBAUM, 2005; GORMAN & CURRAN, 2007). O *network growth* progressivamente adiciona vértices em uma rede, seguindo um determinado critério, e permitindo uma análise da convergência no estado final. Essa abordagem, entretanto, encontra algumas dificuldades no contexto da presente pesquisa. Assumindo que, para simular o processo

evolutivo, adicionássemos vértices de um grafo (denominado ‘fonte’) em outro (denominado ‘alvo’), encontraríamos os seguintes problemas:

1. Como conectar um novo vértice: seria difícil determinar como ligar o novo vértice no grafo ‘alvo’, uma vez que as ligações são determinadas por conta do compartilhamento de significado do verbo respectivo no grupo. Se baseássemos a ligação apenas na topologia do grafo ‘fonte’, a chance de particionar o grafo ‘alvo’ seria grande, pois os outros vértices a que o adicionado se ligaria poderiam não estar presentes (os grafos são pouco similares em termos de conteúdo, como será visto no capítulo 4).
2. Como conectar um vértice existente: seria difícil decidir o que fazer quando chegasse o momento de adicionar um vértice que já existe no grafo ‘alvo’ (um vértice repetido). Se simplesmente adicionarmos novas arestas (como base na topologia do grafo ‘fonte’), a semântica das ligações não seria mais a mesma.
3. Como interpretar os resultados: uma vez que temos vértices de múltiplos grupos etários misturados, ligados segundo múltiplas interpretações (por vezes dos mesmos verbos), o grafo resultante não expressa, de fato, uma evolução, mas a mistura de duas estruturas.

A estratégia proposta na presente pesquisa foi denominada *network involution*, ou involução da rede, e funciona de forma inversa ao *network gorwth*. Pela proposta, os vértices de um grafo são iterativamente excluídos (ao invés de incluídos) de acordo com determinado critério, simulando um processo de involução. Neste caso, não existe um grafo ‘fonte’, uma vez que os vértices excluídos são determinados exclusivamente pelos fatores lingüísticos (os “critérios”). Mesmo que o particionamento possa ainda ocorrer, ele tende a ser uma característica do fim do processo de simulação. Além disso, os dados dos grupos permanecem independentes, o que é a maior vantagem da abordagem baseada em involuções.

As eliminações acontecem sempre em um grafo de indivíduos mais velhos, denominado ‘alterado’, e tomando como referência outro de indivíduos mais novos, denominado ‘objetivo’. Uma vez que são três os conjuntos de dados disponíveis, foram realizadas simulações alterando G2 (tomando G1 como referência) e G3 (tomando G2 como referência). A dinâmica das eliminações baseia-se em dois princípios, extraídos das hipóteses formuladas na seção 3.2:

- Quanto menor a faixa etária dos grupos analisados, maior será a freqüência média dos verbos, maior será a polissemia média e menor será a complexidade sintática média.
- Se o grafo ‘alterado’ for modificado de forma a maximizar a freqüência, maximizar a polissemia e minimizar a complexidade sintática, ele se tornará mais parecido com o grafo ‘objetivo’, tanto em termos de conteúdo quanto de estrutura.

Desses princípios, conclui-se que, ao eliminar os vértices com menor freqüência, menor polissemia e maior complexidade sintática, o grafo ‘alterado’ tende a aumentar sua similaridade em relação ao grafo ‘alvo’.

Em suma, a simulação (figura 3.4) por meio de involuções parte de um grafo e de uma lista ordenada de verbos segundo o fator lingüístico (ou combinação de fatores) em teste. São executados sucessivos ciclos, onde, a cada vez, um vértice é eliminado (com

base na lista ordenada) e são coletadas métricas de teoria dos grafos e teoria dos conjuntos (apresentadas na seção 2.3). Quando o último vértice tiver sido eliminado, o resultado será um histórico de cada métrica coletada a cada ciclo.

Quando empates ocorrem (múltiplos vértices na mesma posição da lista ordenada), um deles é escolhido aleatoriamente para eliminação. No próximo ciclo, a mesma posição será avaliada, até que todos os vértices dos verbos ali posicionados tenham sido eliminados. Para minimizar a influência da aleatoriedade de escolha nos empates, a simulação de cada fator semântico foi realizada 10 vezes, e o histórico final das métricas é resultado da média dessas 10 simulações.

Figura 3.4: Simulação da transformação de G2 para G1 por meio da eliminação ordenada de vértices.

Ainda que a figura 3.4 sugira um procedimento unificado de coleta de métricas, os dados referentes à teoria dos conjuntos e à teoria dos grafos foram coletados em simulações diferentes. A decisão por construir módulos diferentes para os dois tipos de métricas deu-se por questões de cronograma: a implementação da coleta de métricas de teoria dos conjuntos era muito mais simples do que a da teoria dos grafos e pôde ser concluída antes (além de ser utilizada em testes preliminares). De qualquer forma, as diferenças entre os dados dessas duas fontes tende a ser pequena. Afinal, as diferenças podem aparecer apenas quando existem empates e, nestes casos, é feita a média de 10 simulações (em ambos os módulos).

4 RESULTADOS

Este capítulo apresenta os resultados da pesquisa (que foram publicados por GERMANN et al. 2010-a e GERMANN et al. 2010-b). Inicialmente, serão abordadas as relações entre os grafos de cada grupo, bem como uma análise do relacionamento dos fatores lingüísticos entre si e em relação aos grafos. A seguir, serão apresentados os resultados das simulações. Por fim, será feita uma análise dos pontos mais importantes.

4.1 Resultados Iniciais

Comparar os grafos dos três grupos etários permite extrair conclusões preliminares acerca da influência dos fatores lingüísticos na evolução do léxico-mental verbal. A tabela 4.1 consolida estes resultados.

O aumento do vocabulário com a idade fica evidente, representado por *V*: o número de verbos diferentes mencionados pelo grupo em todo o experimento (ou seja, a quantidade de vértices existentes no grafo do grupo). Uma progressiva especialização do vocabulário com a idade também pode ser percebida em diversas métricas. Genericamente, a ‘especificidade’ de uma palavra refere-se ao quão particular é seu significado; é o conceito oposto à ‘generalidade’ (explicada na seção 2.2.2). No contexto desta pesquisa, a especificidade está sendo medida em relação aos vídeos em que o verbo foi mencionado: assumimos que verbos pouco específicos podem ser utilizados para descrever muitos vídeos (por serem mais genéricos), enquanto verbos muito específicos são capazes de descrever poucos vídeos. A repetição média (quantidade média de vídeos em que determinado verbo foi mencionado), em especial, é útil na mensuração dessa grandeza (tabela 4.1). Não apenas a quantidade de repetições é maior quanto mais jovens os integrantes do grupo (portanto menos específico o vocabulário), mas também maior é o desvio padrão (indicando que existem alguns verbos com um alto índice de repetição nos grafos dos indivíduos mais jovens). Em G3, a quantidade de repetições é baixa; os verbos tendem a ser utilizados em poucos vídeos, sugerindo uma grande especificidade no vocabulário. A especialização reflete-se também na quantidade de *hubs* (vértices que foram repetidos; mencionados em mais de um vídeo), que é pequena em G3 e grande em G1 e G2. *D* e $<k>$ apresentam comportamentos diretamente compatíveis com as repetições, decrescendo com o aumento da idade. Uma vez que a repetição leva a um aumento significativo na quantidade de arcos (o verbo repetido liga-se a todos os outros elementos mencionados em cada vídeo), isso se reflete na proporção de arcos por vértices (expressos de formas diferentes por *D* e $<k>$). Em G3, com menos repetições e, portanto, menos ligações entre os vértices, a distância entre eles tende a aumentar, o que resulta em um aumento de *L* no grafo.

As evoluções percebidas em três das propriedades tabela 4.1 não se encontram tão diretamente relacionadas à repetição, ainda que sejam influenciadas por ela de forma

importante. A primeira propriedade é o número de arcos (*E*), que permaneceu praticamente inalterado, por conta também da quantidade de vértices. Quanto maior o número de vértices e maior o número de repetições, maiores tendem a ser os clusters, portanto maior a quantidade de arcos (uma vez que todos os vértices de um cluster ligam-se entre si). Sabendo que, com a idade, a repetição diminui, mas a quantidade de vértices aumenta (e que ambos os fatores influenciam na quantidade de arcos), o que existe é uma compensação, impedindo um aumento no total de arcos. Medidas como *D* e <*k*>, que consideram vértices e arcos na composição do valor, são, dessa forma, mais interessantes no processo de comparação de grafos.

A segunda propriedade é o diâmetro (*diam*). Uma vez que existe um aumento de *V* e uma diminuição da repetição média, seria esperado um aumento de *diam*, o que aconteceu de forma branda. Neste caso, a construção dos grafos foi preponderante para o resultado: uma vez que são inerentemente clusterizados (pela metodologia de construção), é difícil haver um grande incremento. Mesmo a pouca repetição existente acaba sendo suficiente para que todos os caminhos sejam curtos.

Tabela 4.1: Propriedades dos grafos.

Métrica	G1	G2	G3
Número de vértices (*V*)	22	25	31
Número de arcos (*E*)	128	126	126
Caminho mínimo médio (*L*)	1,46	1,6	1,98
Diâmetro (*diam*)	3	3	4
Densidade (*D*)	0,55	0,42	0,27
Coeficiente de clusterização (*C/s*)	0,84	0,78	0,78
Conectividade média (<*k*>)	Méd.: 11,64 DP: 6,73	Méd.: 10,08 DP: 4,86	Méd.: 8,13 DP: 4,76
Repetição média	Méd.: 5,23 DP: 4,41	Méd.: 3,76 DP: 3,15	Méd.: 2,19 DP: 1,58
Porcentagem de *hubs*	77,27%	80%	51,61%

A terceira propriedade é o coeficiente de clusterização. Dentro de um cluster, todos os vértices possuem *C/s* = 1 (inicialmente, os três grafos são conexos, portanto não existe divisão por *S*), com exceção daqueles que tiverem sido mencionados em múltiplos vídeos. Como estes terão vizinhos em dois (ou mais) clusters, é natural que a quantidade de arcos não presentes entre os vizinhos seja grande. Dessa forma, *C/s* irá aproximar-se de 1 em duas situações: se a quantidade de repetições for muito pequena (existem poucos vértices com vizinhos em múltiplos clusters) ou se a quantidade de repetições for muito grande (muitos vértices passam a ser vizinhos em múltiplos clusters, e o grafo aproxima-se da completude). Em G1, existe uma grande quantidade de *hubs* e uma grande média de repetições, ou seja, uma grande quantidade de vértices que repete parte de sua vizinhança em outro cluster, elevando *C/s*. G2 possui ainda mais *hubs*, mas possui menor grau de repetição, ou seja, muitos *hubs* que poucas vezes repetem parte da vizinhança, diminuindo *C/s*. Por fim, G3 possui uma quantidade menor de *hubs*, mas uma média de repetições também menor. Isso significa que existem muitos vértices com *C/s* = 1, por estarem em apenas um cluster, mas que os *hubs*

tendem a apresentar um baixo *C/s*, por terem muito pouca vizinhança compartilhada, diminuindo a média global.

O efeito da repetição fica evidenciado ao verificarmos a evolução das repetições nos três grafos (figura 4.1). Os verbos de G3 foram mencionados em, no máximo, 7 videoclipes diferentes, de G2 em 11 e de G1 em 14. Além disso, G1 possui cinco verbos (quase 20% do total) com 11 ou mais repetições (de um máximo possível de 17).

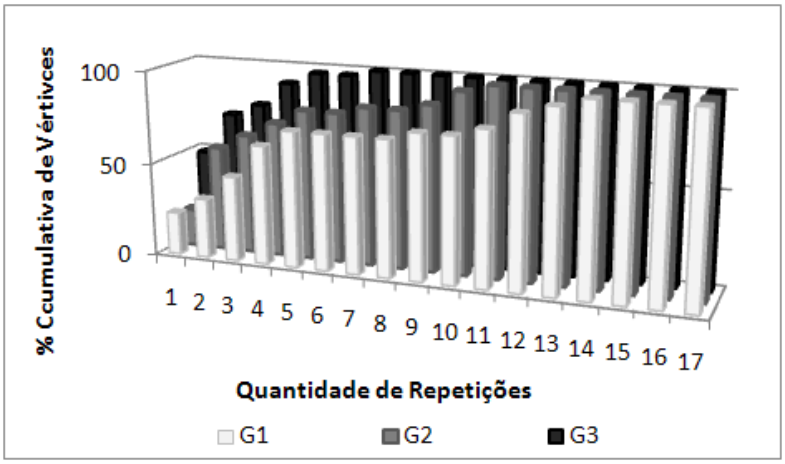


Figura 4.1: Histograma cumulativo da porcentagem de verbos com no máximo determinado número de repetições.

Quanto ao conteúdo, os grafos compartilham uma quantidade substancial de verbos: são 12 verbos (de um total de 44) comuns aos três conjuntos. Uma análise da similaridade dos conjuntos é apresentada na tabela 4.2. A primeira linha apresenta a quantidade bruta de verbos comuns aos grafos sendo analisados em cada coluna. A segunda linha apresenta a porcentagem de verbos mencionados pelo grupo mais velho que não foram mencionados no grupo mais novo, enquanto a terceira linha apresenta a mesma informação, mas em relação ao grupo mais jovem. A quarta linha, por fim, apresenta o coeficiente de Jaccard, que é diretamente influenciado pelas medidas das linhas anteriores.

Tabela 4.2: Comparações entre os grafos[17].

Critério de Comparação	G1-G2	G2-G3	G1-G3	Todos
Verbos comuns aos grafos	16	17	13	12
Verbos presentes no grafo do grupo mais velho, não presentes no grafo do grupo mais novo (%)	36	45,16	58,06	-
Verbos presentes no grafo do grupo mais novo, não presentes no grafo do grupo mais velho (%)	27,27	32	40,9	-
Coeficiente de Jaccard	0,52	0,44	0,33	-

Analisando a tabela 4.2, é possível verificar uma similaridade maior entre os conteúdos de G2 e de G1 do que entre os de G2 e de G3, coincidindo com a

[17] As únicas comparações relevantes são entre os pares G1-G2 e G2-G3. Os valores de G1-G3 foram apresentados apenas para referência: não foi feita uma simulação entre estes conjuntos.

proximidade das idades médias de cada grupo: a diferença entre G1 e G2 é de apenas 2 anos, enquanto entre G2 e G3 é de mais de 15. A tabela 4.1 indica que o mesmo se verifica na estrutura dos grafos, com os valores de G2 apresentando maior similaridade em relação a G1 do que em relação a G3 em 6 dos 9 resultados apresentados.

As hipóteses formuladas no capítulo 3 encontram respaldo estatístico na medição dos fatores lingüísticos sobre os três grafos dos grupos de indivíduos (tabela 4.3[18]). Os resultados com maior polissemia, maior freqüência e menor complexidade sintática foram encontrados, principalmente, em G1, e o oposto em G3, com G2 permanecendo no meio. Exceções foram encontradas (em relação a G2) nos valores associados aos escores PolWord e FreqYahoo.

Tabela 4.3: Média e desvio padrão dos escores lingüísticos. Calculada sobre os valores associados aos vértices dos grafos (*type*)[19].

Escore	G1	G2	G3	Total
PolWord	Méd.: 10,55 DP: 8,44	Méd.: 10,64 DP: 8,17	Méd.: 10,48 DP: 10,65	Méd.: 10,14 DP: 9,66
PolHou	Méd.: 21,59 DP: 18,89	Méd.: 20,84 DP: 18,11	Méd.: 16,26 DP: 16,06	Méd.: 17,32 DP: 17,25
FreqYahoo	Méd.: 15441904 DP: 46290242	Méd.: 18443193 DP: 49620807	Méd.: 10419263 DP: 28033898	Méd.: 13816222 DP: 39701981
FreqFlorian	Méd.: 44,05 DP: 104,51	Méd.: 35,92 DP: 99,37	Méd.: 17,84 DP: 47,36	Méd.: 27,77 DP: 82,23
CompSint	Méd.: 8,32 DP: 9,95	Méd.: 10,52 DP: 10,08	Méd.: 13,16 DP: 9,39	Méd.: 11,55 DP: 9,82

A existência de exceções deve-se ao fato de as medições avaliarem apenas o conteúdo dos grafos, desconsiderando a importância relativa de cada verbo dentro do vocabulário. Dessa forma, um verbo mencionado por apenas duas pessoas (o suficiente para não ser excluído na 'limpeza' mencionada no capítulo 3) e outro mencionado por todo o grupo contribuem igualmente para o resultado. Uma vez que é grande a variação de alguns escores dentro dos grupos (apêndice D), e que a quantidade de verbos é relativamente pequena, os resultados acabam distorcidos.

A fim de verificar o impacto da ponderação (quantidade de vezes em que o verbo é mencionado), essa análise foi repetida diretamente sobre as respostas dos indivíduos. A tabela 4.4 apresenta as médias coletadas sobre todas as respostas de cada grupo:

[18] Por questões de organização e espaço, a porção decimal do escore FreqYahoo foi suprimida nas tabelas 4.3 e 4.4.

[19] A coluna "Total" mostra médias e desvios padrão sobre todos os verbos mencionados nos três grafos, sendo cada verbo considerado apenas uma vez (independentemente de em quantos grafos ele estava presente). Dessa forma, é possível compreender como a média desta coluna, quando referente ao escore PolWord, pode ser menor do que a média dos grafos em separado: os verbos com valores maiores foram mencionados em múltiplos grafos. Assim, ao unir os três grafos, havia menos verbos com escores maiores e mais verbos com escores menores, o que acabou elevando as médias locais, mas diminuindo a média global.

primeiramente foi feita a média por indivíduo (considerando apenas as respostas válidas), e, por fim, uma média destes resultados intermediários (resultando no valor do grupo). Neste caso, todos os fatores lingüísticos confirmaram as hipóteses a eles associadas: com o aumento da idade, há uma diminuição da freqüência e polissemia, juntamente com um aumento na complexidade sintática.

As diferenças entre as tabelas 4.3 e 4.4 reforçam a importância dos fatores lingüísticos. Por exemplo, ainda que o escore PolWord seja similar nos três grafos (analisando a tabela 4.3), os verbos com maior polissemia foram muito mais utilizados nas respostas de G1 do que de G2 e G3 (analisando a tabela 4.4.). Conclusões análogas aplicam-se aos demais escores.

Tabela 4.4: Média e desvio padrão dos escores lingüísticos. Calculado sobre os valores associados às respostas de cada indivíduo (*token*).

Escore	G1	G2	G3	Total
PolWord	Méd.: 16,25 DP: 3,86	Méd.: 14,66 DP: 2,51	Méd.: 11,13 DP: 2,35	Méd.: 14,01 DP: 3,66
PolHou	Méd.: 26,93 DP: 6,79	Méd.: 23,02 DP: 4,49	Méd.: 17,82 DP: 3,28	Méd.: 22,59 DP: 6,27
FreqYahoo	Méd.: 10788194 DP: 5563701	Méd.: 9277047 DP: 6984359	Méd.: 8927866 DP: 5816293	Méd.: 9664369 DP: 6168407
FreqFlorian	Méd.: 43,44 DP: 17,04	Méd.: 35,71 DP: 15,16	Méd.: 21,22 DP: 9,08	Méd.: 33,46 DP: 16,85
CompSint	Méd.: 4,75 DP: 3,2	Méd.: 6,61 DP: 1,85	Méd.: 10,7 DP: 1,96	Méd.: 7,35 DP: 3,46

Visando confirmar se as diferenças de médias encontradas na tabela 4.4 são ou não significativas, foram realizadas comparações estatísticas sobre os escores de cada resposta (cada escore foi testado independentemente). Primeiramente, foi verificado se as variâncias dos três grupos (G1, G2 e G3) apresentavam diferenças significativas. O teste de Levene não encontrou heterogeneidade: mesmo nos casos em que o teste apresentou diferença significativa (PolWord, PolHou, FreqFlorian e CompSint), foi possível continuar a análise, uma vez que há uma tolerância quando o valor da razão entre a maior e a menor variância é menor do que 7 (o que de fato aconteceu). O passo seguinte foi a ANOVA (*analisys of variance*, ou, análise da variância), que encontrou diferenças significativas entre as médias dos grupos para o escores PolWord, PolHou, FreqFlorian e CompSint. Nestes casos, o teste de Tukey foi utilizado para comparar as médias dos grupos (dois a dois), mostrando que elas diferem significativamente entre si. Quanto ao escore FreqYahoo, a ANOVA não mostrou uma diferença significativa entre as médias (p>0,05), portanto não há uma diferença significativa entre os grupos. Em suma, a não ser por FreqYahoo, podemos afirmar que existe forte indício estatístico apontando para uma diminuição da freqüência, diminuição da polissemia e aumento da complexidade sintática nos verbos que compõem o vocabulário dos grupos, com o aumento da idade de seus participantes. As tabelas dos resultados parciais apresentados pelo software SPSS podem ser vistas no anexo A.

A análise estatística também foi utilizada para comparar o relacionamento entre os escores lingüísticos. As hipóteses propostas no capítulo 3 sugerem que os três tipos de fatores lingüísticos influenciam no processo de aquisição de verbos. Em vista disso, é de se esperar que a ordem de eliminação resultante de todos seja similar. Assim, foram realizados testes de correlação utilizando os coeficientes de Kendall (τ) e Spearman (ρ) sobre as listas ordenadas pelos fatores lingüísticos. Ambas as medidas variam entre -1 e 1, sendo que um coeficiente positivo significa uma correlação direta (os valores de uma lista tendem a aumentar quando os valores da outra aumentam) enquanto valores negativos indicam uma correlação inversa (quando os valores de uma aumentam, os da outra diminuem). Um coeficiente de 0 indica ausência de correlação, ou seja, o valores de uma lista variam independentemente dos valores da outra. Os resultados (tabela 4.5) indicam a existência de correlação (p = 0,01) entre todos os fatores. O escore que apresentou menor grau de correlação com os outros foi CompSint, seguido por FreqFlorian. Os dois fatores tiveram como fonte o corpus Florianópolis, que, como mencionado no capítulo 3, não apresentou resultados para 25 dos 44 verbos procurados, que acabaram, por conseguinte, posicionados na mesma posição em ambas as listas. Em vista disso, o maior grau de correlação foi encontrado justamente entre esses dois escores. O desempenho pior da complexidade sintática deve-se à baixa quantidade de classificações possíveis: apenas cinco (sendo uma reservada exclusivamente aos verbos não encontrados no corpus).

Tabela 4.5: Coeficiente de correlação de Kendall (τ) e de Spearman (ρ) entre todos os fatores lingüísticos.

Escore	PolWord	PolHou	FreqYahoo	FreqFlorian	CompSint
PolWord	τ: 1,0 ρ: 1,0	τ: 0,637 ρ: 0,800	τ: 0,637 ρ: 0,795	τ: 0,520 ρ: 0,641	τ: 0,408 ρ: 0,527
PolHou	τ: 0,637 ρ:0,800	τ: 1,0 ρ: 1,0	τ: 0,611 ρ: 0,800	τ: 0,573 ρ: 0,700	τ: 0,505 ρ: 0,623
FreqYahoo	τ: 0,637 ρ: 0,795	τ: 0,611 ρ: 0,800	τ: 1,0 ρ: 1,0	τ: 0,507 ρ: 0,660	τ: 0,462 ρ: 0,595
FreqFlorian	τ: 0,520 ρ: 0,641	τ: 0,573 ρ: 0,700	τ: 0,507 ρ: 0,660	τ: 1,0 ρ: 1,0	τ: 0,807 ρ: 0,925
CompSint	τ: 0,408 ρ: 0,527	τ: 0,505 ρ: 0,623	τ: 0,462 ρ: 0,595	τ: 0,807 ρ: 0,925	τ: 1,0 ρ: 1,0

4.2 Resultados da Simulação

A presente pesquisa visa apontar a relação entre os fatores lingüísticos e a facilidade de aprendizado. Ainda que os resultados apresentados na seção anterior sugiram a confirmação das hipóteses formuladas no capítulo 3, verificaremos, nesta seção, o impacto direto dos fatores lingüísticos no processo evolutivo. Serão apresentados, portanto, os resultados das simulações de acordo com as listas ordenadas pelos escores lingüísticos.

De acordo com os princípios da seção 3.3, ao eliminarmos, do grafo ‘alterado’ (dos indivíduos com maior idade média), os verbos com menor freqüência, menor polissemia e maior complexidade sintática, sua estrutura e conteúdo devem tornar-se mais

parecidos com os do grafo ‘objetivo’ (dos indivíduos com menor idade média). Por conta disso, ainda que os gráficos das simulações apresentem todo o histórico de eliminação dos verbos, o início é a parte mais importante. Afinal, uma vez que a eliminação dá-se em ordem ascendente de polissemia e freqüência, e descendente de complexidade sintática, são os primeiros verbos eliminados aqueles que supomos estarem diferenciando os dois grafos.

Para facilitar a interpretação dos resultados, os gráficos contam com linhas referentes à eliminação randômica. Cada ponto da linha é resultado da média de 10 execuções completas utilizando a eliminação aleatória.

4.2.1 Resultados das Métricas de Teoria dos Grafos

Os resultados serão apresentados agrupados por métrica, e com as duas simulações lado a lado. Os gráficos da esquerda correspondem ao grafo G2 sendo alterado para aproximar-se de G1, enquanto a coluna da direita mostra G3 sendo alterado para aproximar-se de G2. As eliminações segundo cada um dos escores são mostradas uma abaixo da outra. Cada gráfico é composto por duas linhas principais: (a) a de simulação (utilizando o método *network involution*) segundo cada fator lingüístico ou combinação (denominada “Simulação”) e (b) a de eliminação utilizando o critério randômico (denominada “Randômico”). Adicionalmente, cada figura apresenta a medição da métrica no grafo ‘objetivo’ (uma linha reta, tracejada, mais grossa, denominada “Referência”).

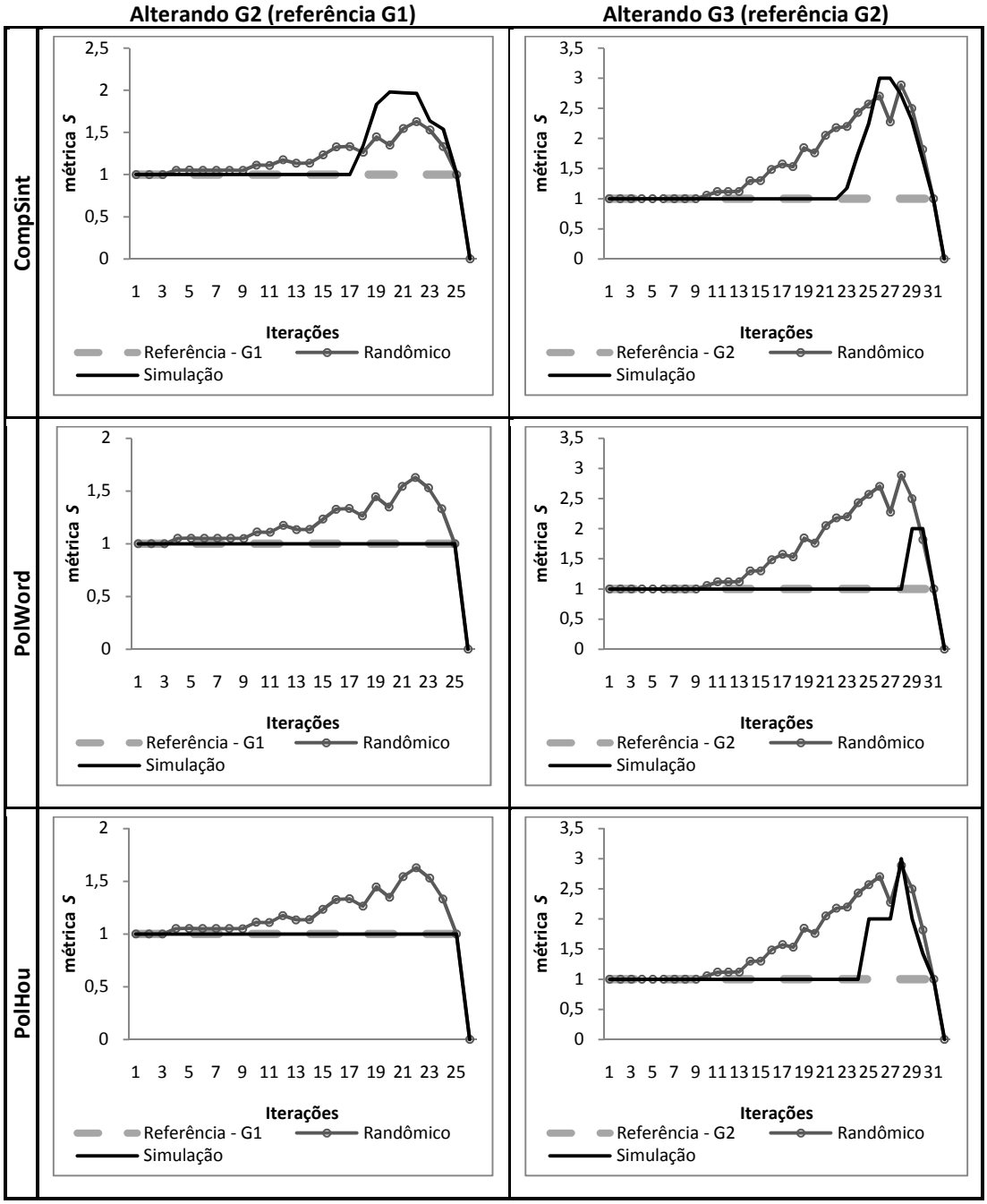
O primeiro fator a ser analisado é o número de subgrafos (*S*), ou seja, a quantidade de subgrafos conexos existentes em um dado momento (figura 4.2). Os subgrafos iniciais (criados a partir dos vídeos) foram unidos por compartilharem parte do significado. Se o vocabulário não é uno (i.e., o grafo não é conexo), então o compartilhamento de significado é baixo. Em outras palavras, o particionamento de um grafo remete a uma heterogeneidade no vocabulário.

Nos gráficos, é possível perceber que, em ambas as simulações, a eliminação randômica particionou o grafo já nas primeiras iterações (inicialmente, existe sempre um único subgrafo conexo, que corresponde ao grafo inteiro). O fato de as eliminações por meio de fatores lingüísticos terem gerado múltiplos subgrafos apenas no final (ou nem ao menos terem gerado), demonstra qualidade nos resultados ao manter estável a estrutura inicial. Destacam-se os resultados do escore CompSint, que, mesmo classificando os verbos em poucos grupos– e randomizando os resultados internamente a cada um –, apresentou um comportamento similar ao dos demais fatores.

A simulação de G3 foi muito mais afetada pelas eliminações do que a de G2, gerando uma quantidade maior de partições (um pico perto de três partições) em todos os fatores lingüísticos. A principal justificativa para esse resultado é a baixa quantidade de *hubs* (verbos que ligam os clusters uns aos outros). Se a quantidade de caminhos entre os clusters dos vídeos (que foram submetidos ao *merge*) é pequena, a chance de isolamento aumenta.

Nos resultados apresentados, é possível perceber variações menores do que 1, o que não seria esperado imediatamente. A explicação para isso está na metodologia: uma vez que os gráficos são resultado da média de 10 simulações, os pontos onde existe empate terão variações mais tênues. Uma vez que a eliminação randômica considera um empate de todos os vértices, é normal que variações pequenas sejam perceptíveis desde o início da simulação. Particularmente em relação à métrica *S*, um valor maior do que 0 nos

pontos iniciais demonstra que o particionamento do grafo é possível mesmo com poucas eliminações. Essa evidência reforça a importância dos resultados percebidos com os fatores lingüísticos, onde o particionamento ocorre apenas no final.



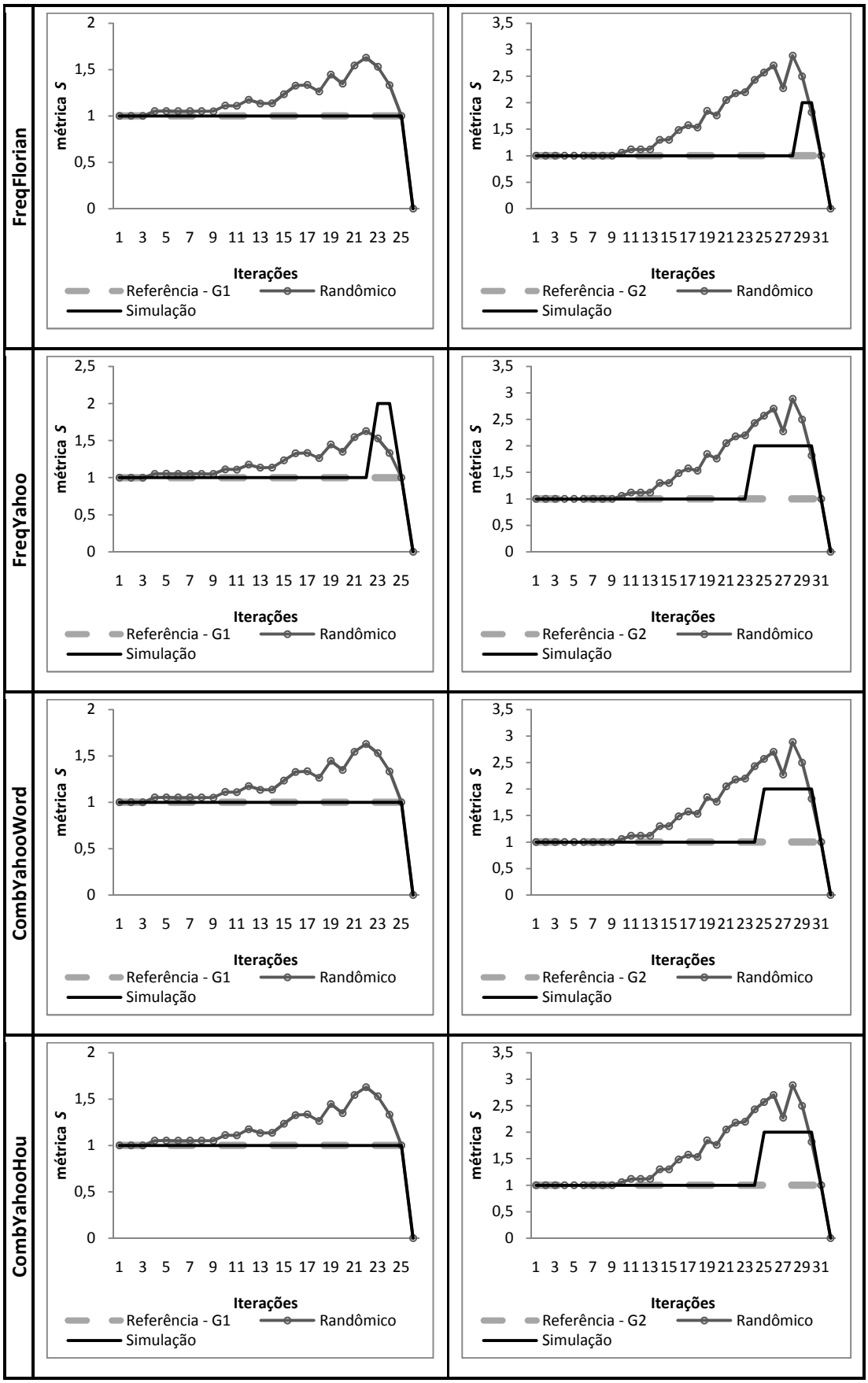
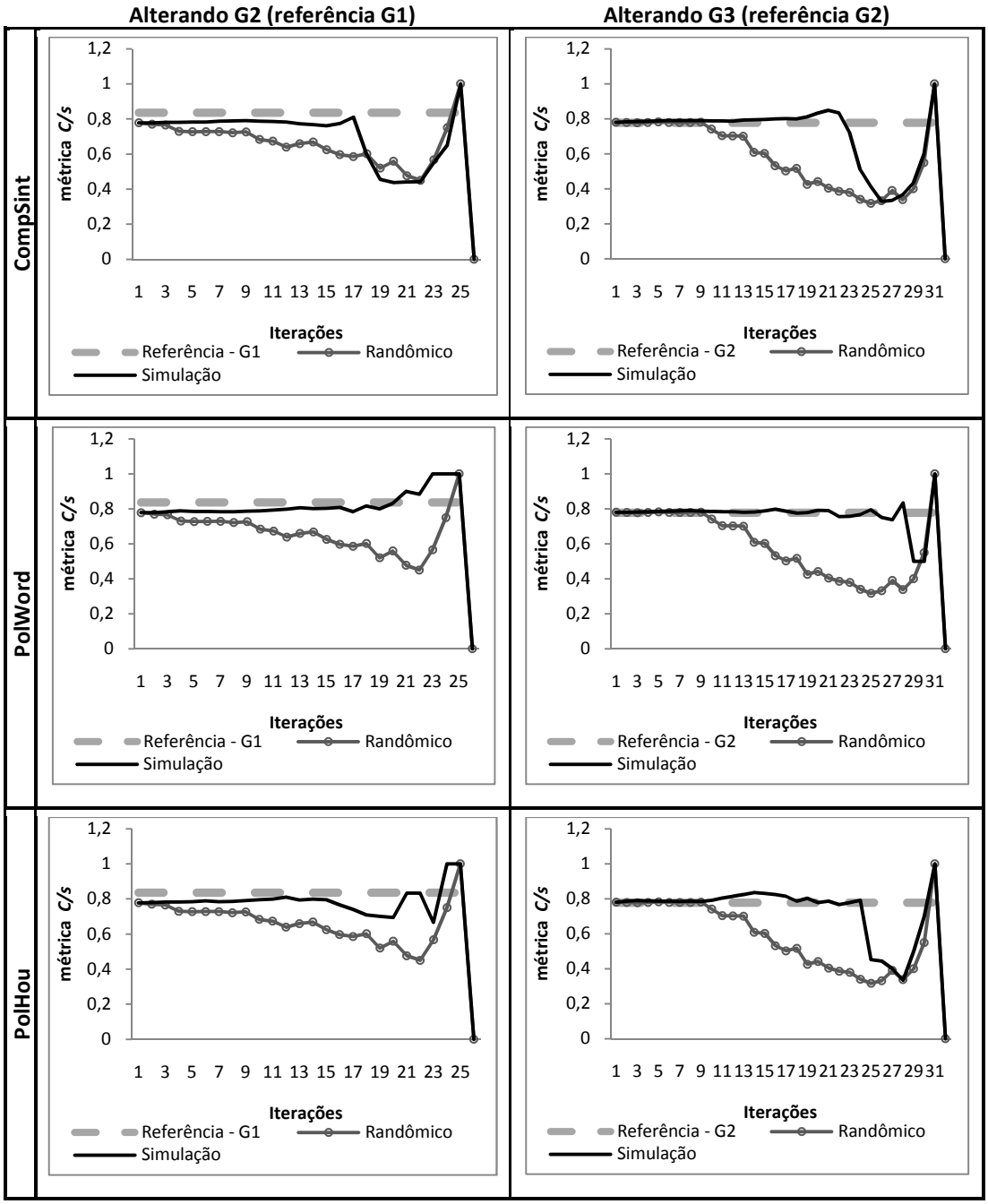


Figura 4.2: Evolução da métrica ‘número de subgrafos’ (*S*) nas duas simulações.

A métrica original ‘coeficiente de clusterização’ (*C*), quando testada, apresentou uma baixa variação em todos os gráficos, tanto pela eliminação randômica quanto por fatores lingüísticos. Esse resultado está de acordo com o esperado: uma vez que os grafos foram construídos de forma clusterizada, a eliminação de vértices tende a não afetar globalmente a estrutura, apenas diminuindo os clusters até extinguí-los. Ao combinar o coeficiente de clusterização com o número de subgrafos, criando a métrica *C/s* (como descrito na seção 2.3.1), conseguiu-se uma medição mais adequada a grafos desconexos (figura 4.3). A eliminação randômica foi novamente bastante afetada, uma vez que gera subgrafos desde o início da simulação.



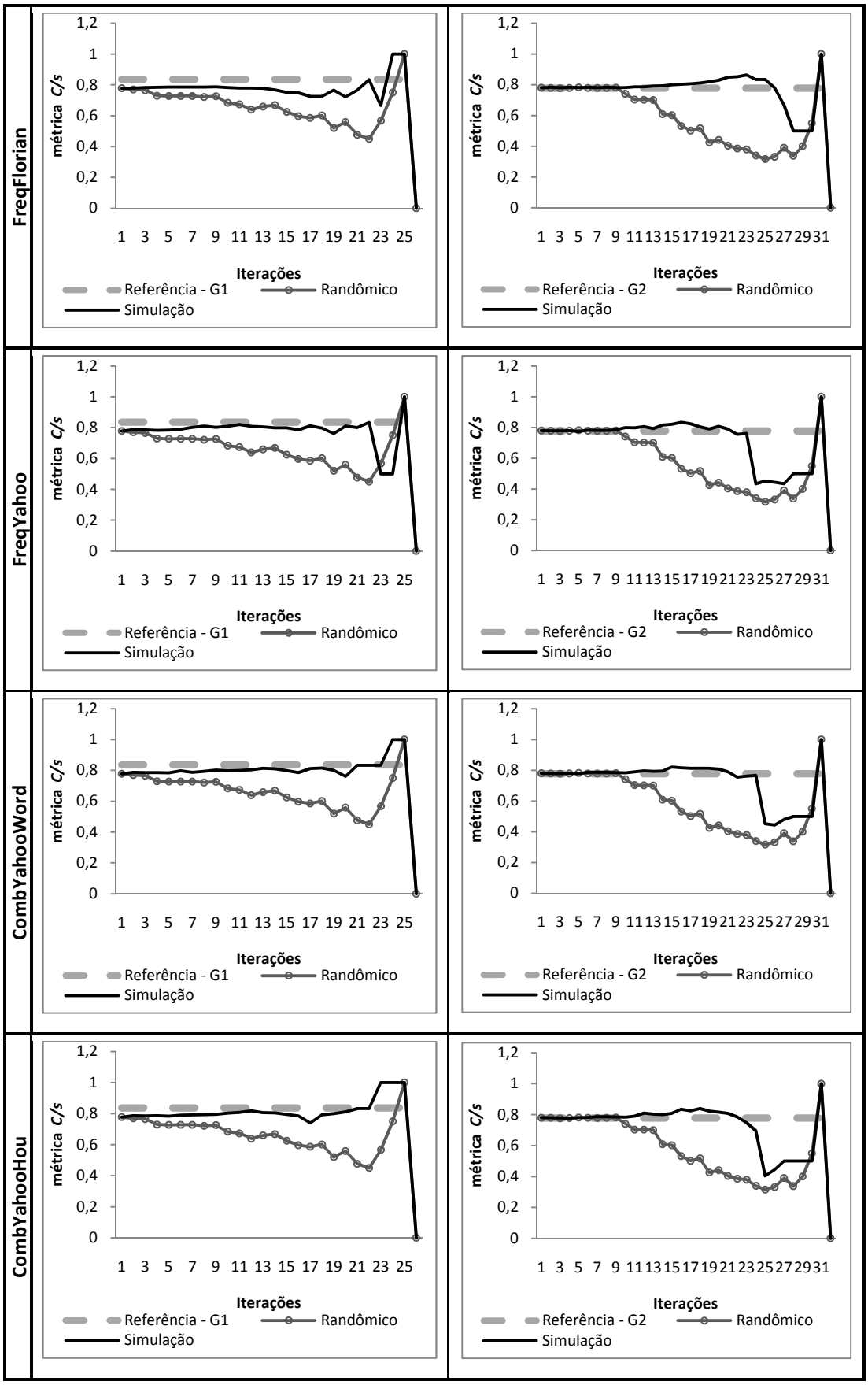
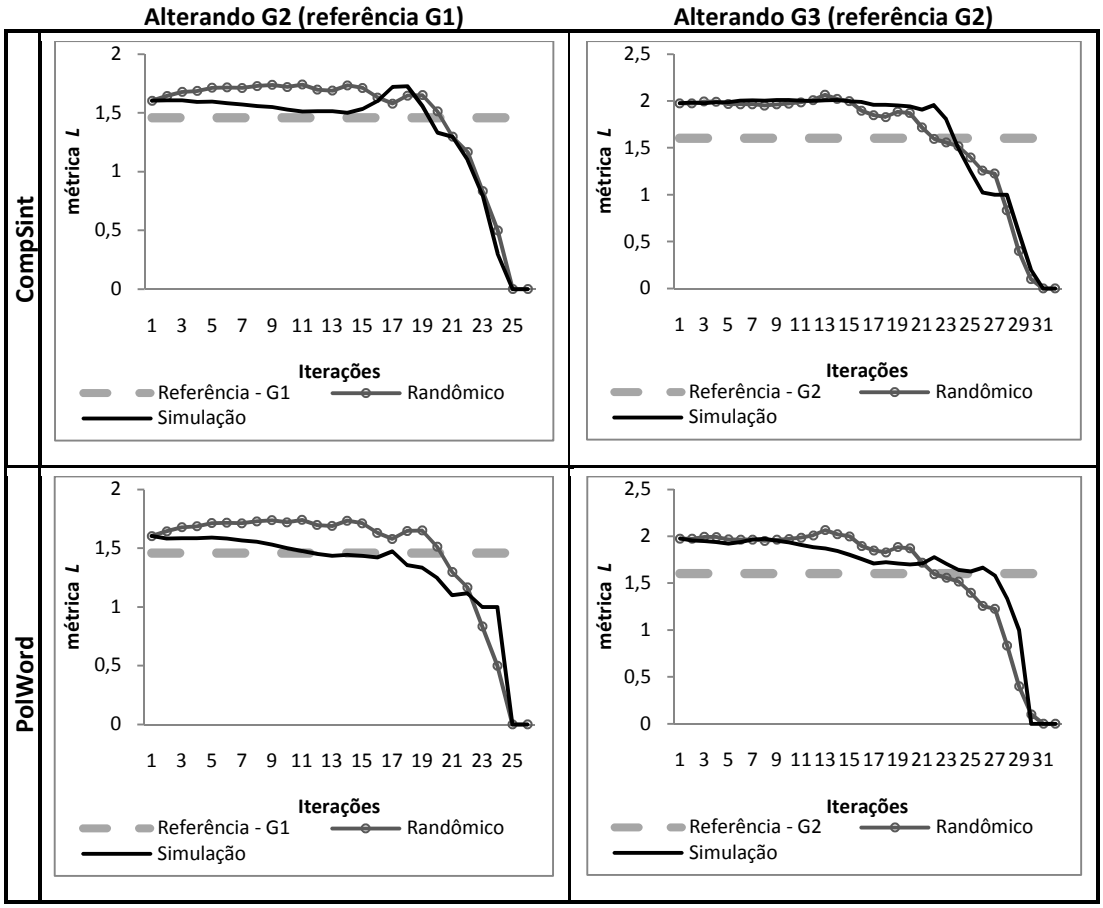
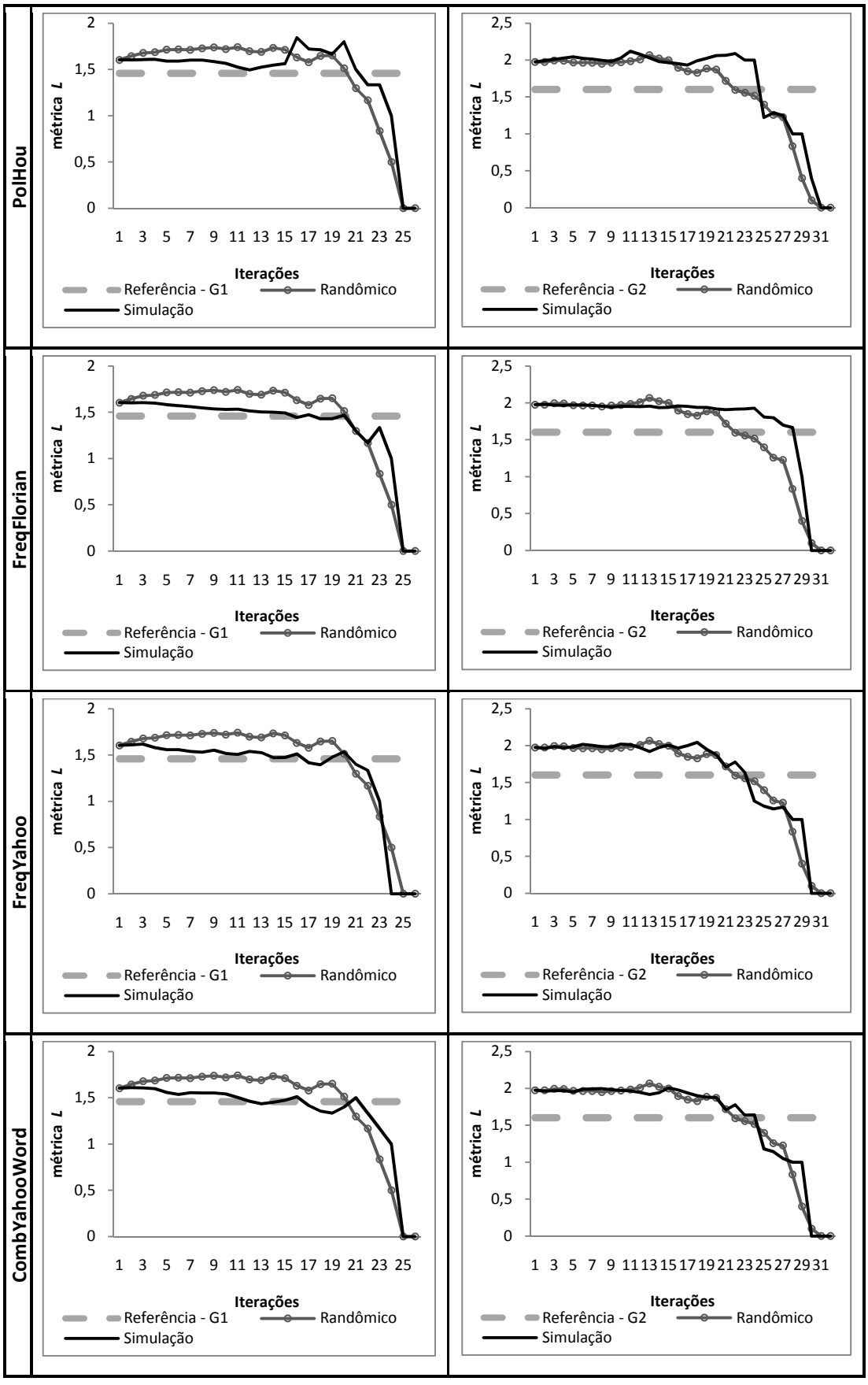


Figura 4.3: Evolução da métrica ‘coeficiente de clusterização com subgrafos’ (*C/s*) nas duas simulações.

O caminho mínimo médio (*L*) apresentou uma ótima evolução na primeira simulação, com os fatores lingüísticos tendendo à linha de referência desde o início e piorando apenas no final (figura 4.4). Entretanto, a segunda simulação teve uma evolução quase randômica. O principal motivo para este resultado é a estrutura inicial dos grafos (em especial a baixa densidade e a baixa quantidade de *hubs* de G3). Os grafos tornam-se progressivamente mais esparsos com o aumento da idade dos indivíduos, mas a diferença entre G3 e G2 é maior do que a entre G2 e G1: a quantidade de vértices em G3 é proporcionalmente muito grande em relação à quantidade de arestas (bastante evidente na figura 3.2). Assim, mesmo com uma eliminação ideal de vértices (buscando diminuir *L*), seriam necessárias muitas exclusões até que o efeito fosse perceptível. O resultado é a impossibilidade de aproximar adequadamente a linha de referência.




PolHou
FreqFlorian
FreqYahoo
CombYahooWord
métrica L
Iterações
Referência - G1
Referência - G2
Randômico
Simulação

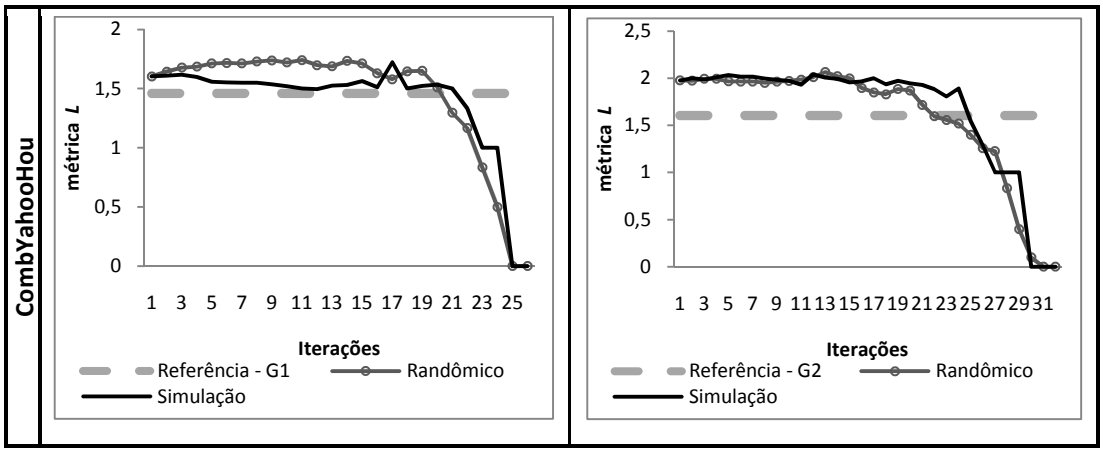
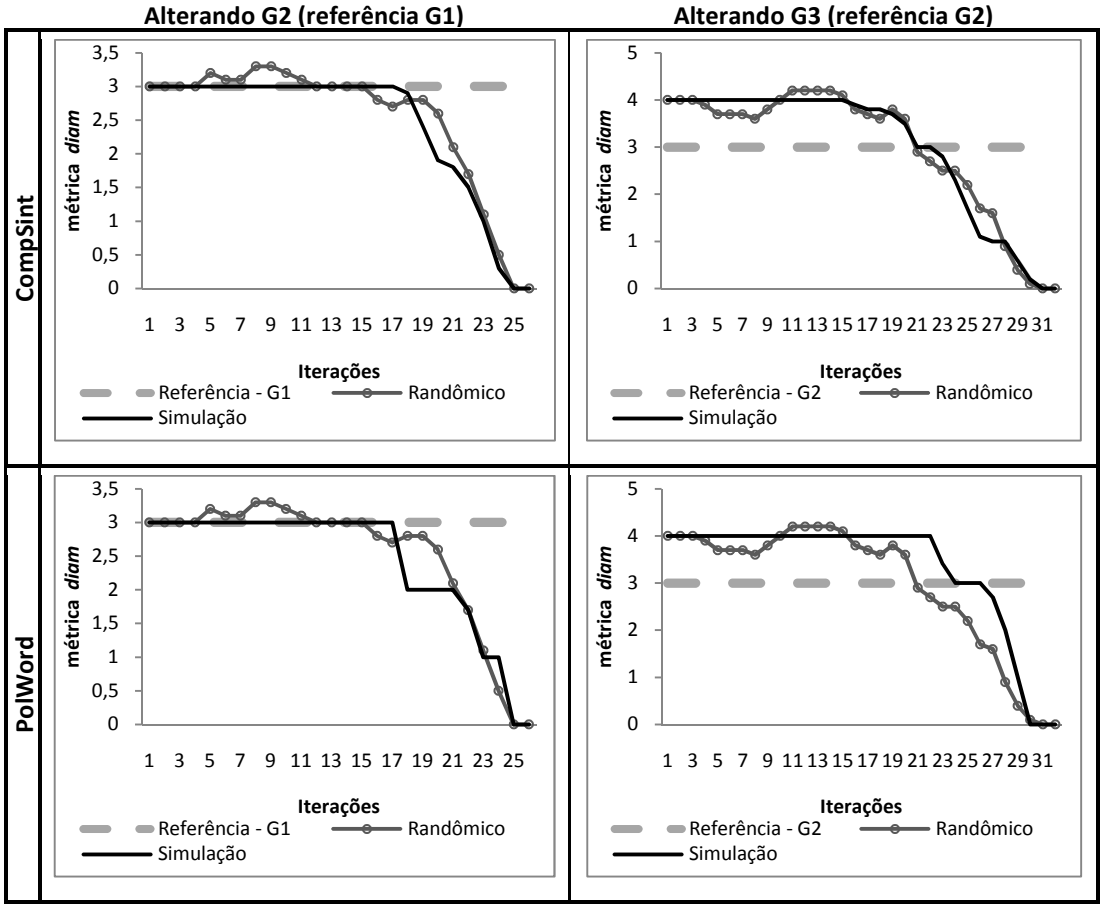
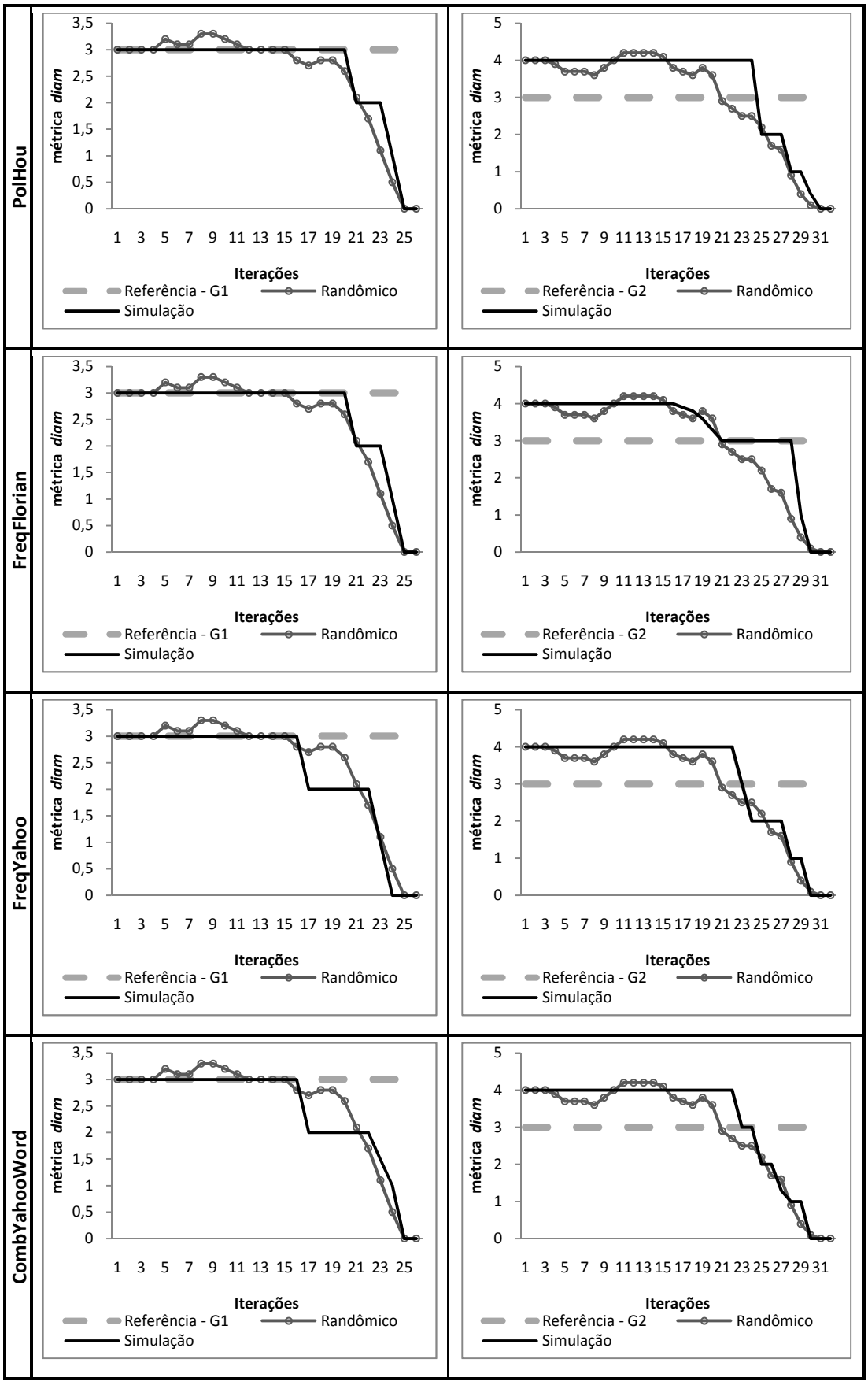


Figura 4.4: Evolução da métrica ‘caminho mínimo médio’ (*L*) nas duas simulações.

Em ambas as simulações, a evolução do diâmetro (*diam*) é similar à da linha randômica, com a diferença de que esta variou muito enquanto aquela apresentou maior estabilidade, com linhas mais suaves (figura 4.5). Na primeira simulação, o resultado é evidentemente positivo, uma vez que a evolução dos fatores lingüísticos coincide perfeitamente com a linha de referência. A segunda simulação, entretanto, é de difícil interpretação, uma vez que os valores iniciais são diferentes. De qualquer forma, a estabilidade do diâmetro é um resultado satisfatório.

Os resultados de *diam*, assim como os de *S*, demonstram claramente o efeito dos empates. É possível perceber, novamente, uma variação menor do que 1, devido a empates, em diversos momentos, principalmente na eliminação randômica.




PolHou
FreqFlorian
FreqYahoo
CombYahooWord
métrica diam
Iterações
Referência - G1
Referência - G2
Randômico
Simulação

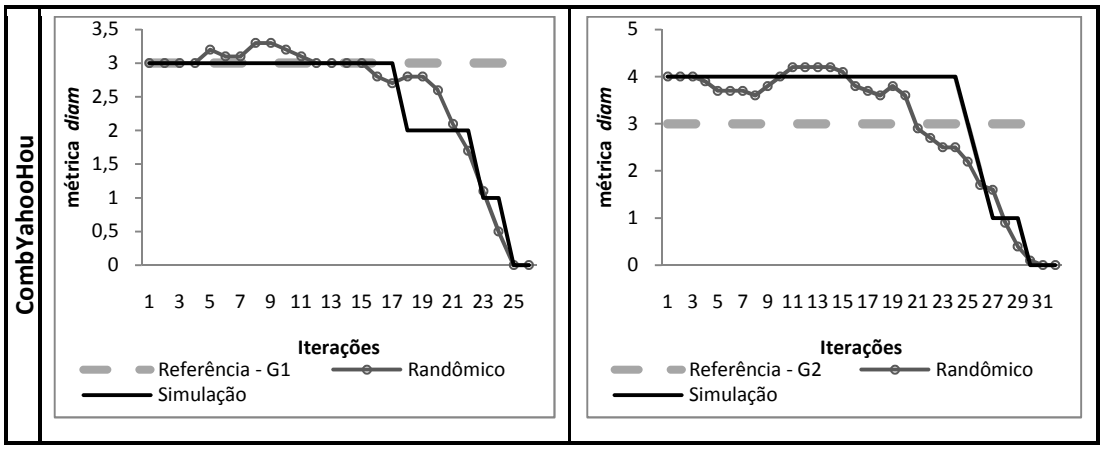
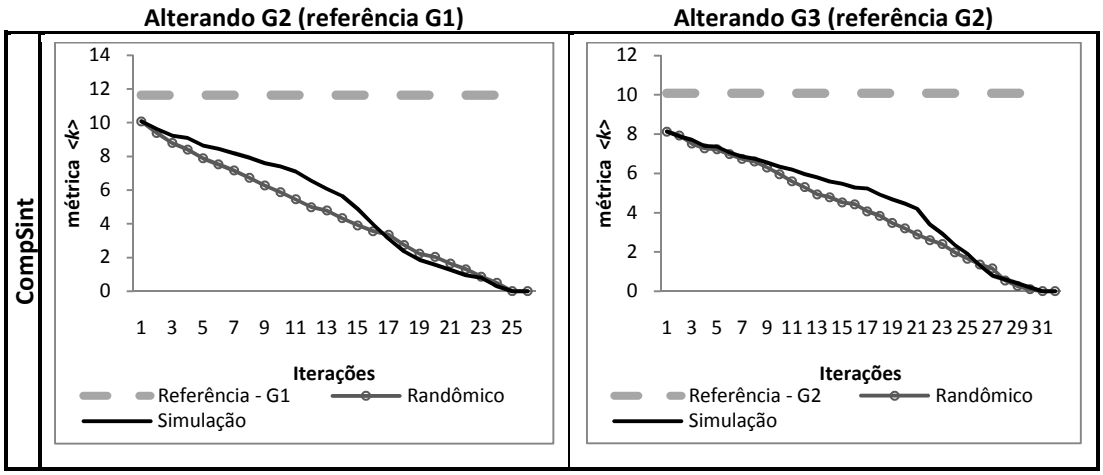


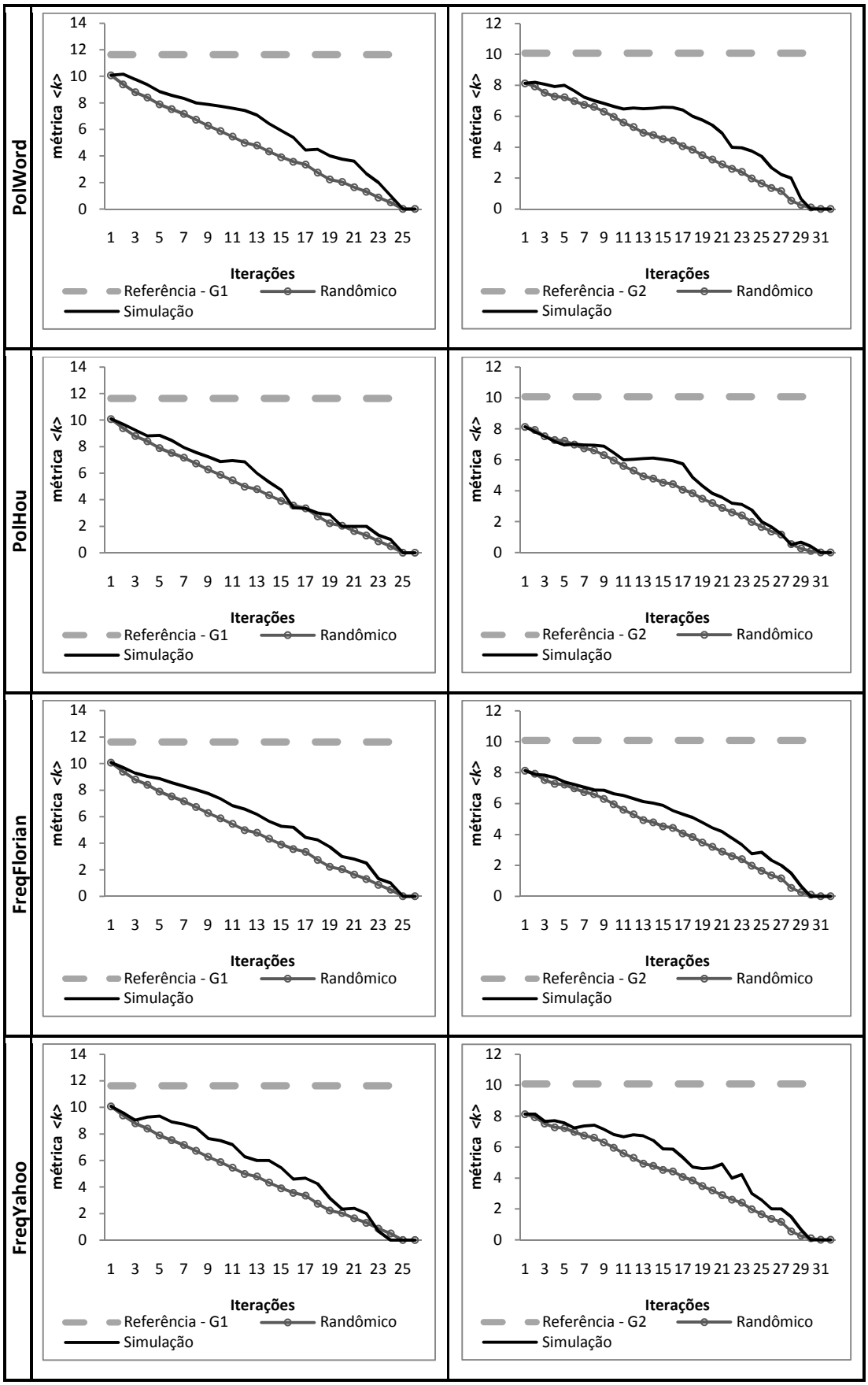
Figura 4.5: Evolução da métrica 'diâmetro' (*diam*) nas duas simulações.

A evolução dos fatores lingüísticos em relação à conectividade média (<*k*>) foi satisfatória em ambas as simulações: as linhas permaneceram acima da medição randômica, ao menos no início (figura 4.6). Ainda que as primeiras iterações diferenciem-se pouco do resultado randômico, os maiores incrementos aconteceram, na maior parte das vezes, na metade inicial dos gráficos, indicando que os verbos eliminados no início eram, de fato, os que diferenciavam as duas estruturas.

Estamos dando maior importância à localização do incremento no gráfico, por este representar uma real aproximação dos dois grafos comparados. Se verificarmos a métrica PolWord, por exemplo, será possível perceber que, na metade final do gráfico, a linha "Simulação" encontra-se mais distante da linha "Randômico" do que na metade inicial. Entretanto, essa relação entre as duas linhas constitui apenas uma manutenção do distanciamento ocorrido na primeira metade: as linhas "Simulação" e "Randômico" tornam-se quase paralelas, antes de se aproximarem novamente no final. Esse paralelismo indica que as duas eliminações estão modificando o grafo de forma similar, e não um distanciamento de fato (que ocorreu apenas na metade inicial).

A presença de uma zona quase randômica nas primeiras iterações é explicada pela estrutura inicial dos grafos. Para que haja um aumento de <*k*>, os nodos com menos ligações devem ser eliminados, permanecendo um núcleo mais fortemente conectado. A criação desse núcleo pode demandar a eliminação de muitos nodos, principalmente de G3, pelos mesmos motivos apresentados quando analisando a métrica *L* nesta seção: baixa quantidade de *hubs* e baixa densidade.




PolWord
PolHou
FreqFlorian
FreqYahoo
métrica <k>
Iterações
Referência - G1
Referência - G2
Randômico
Simulação

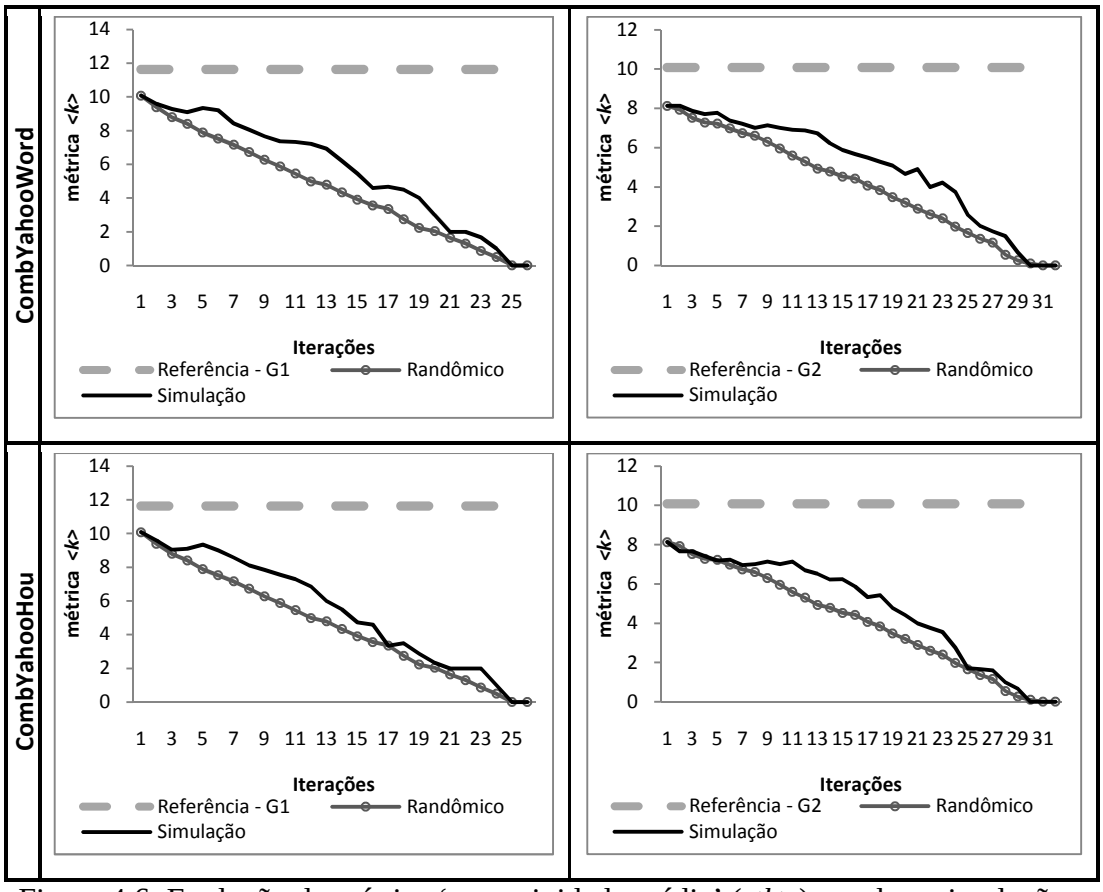
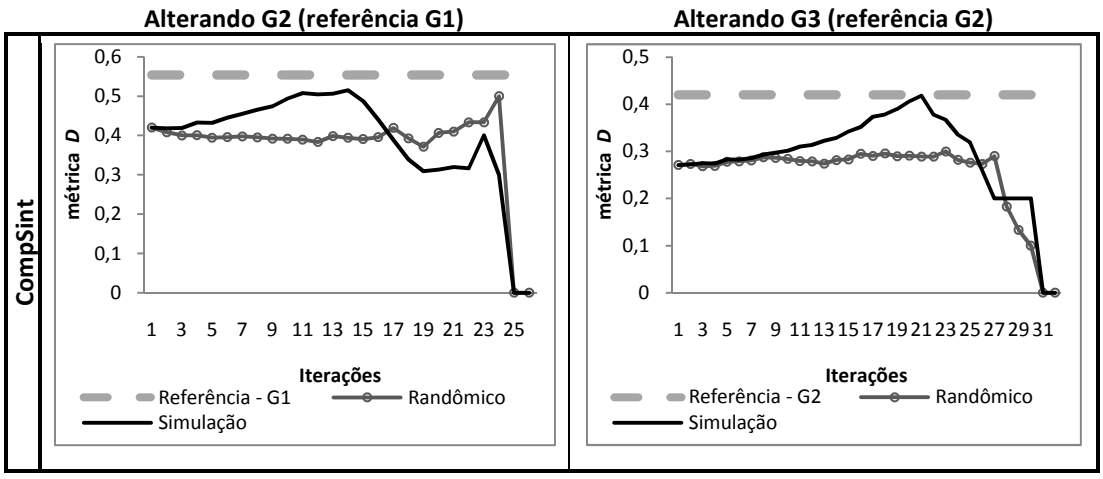
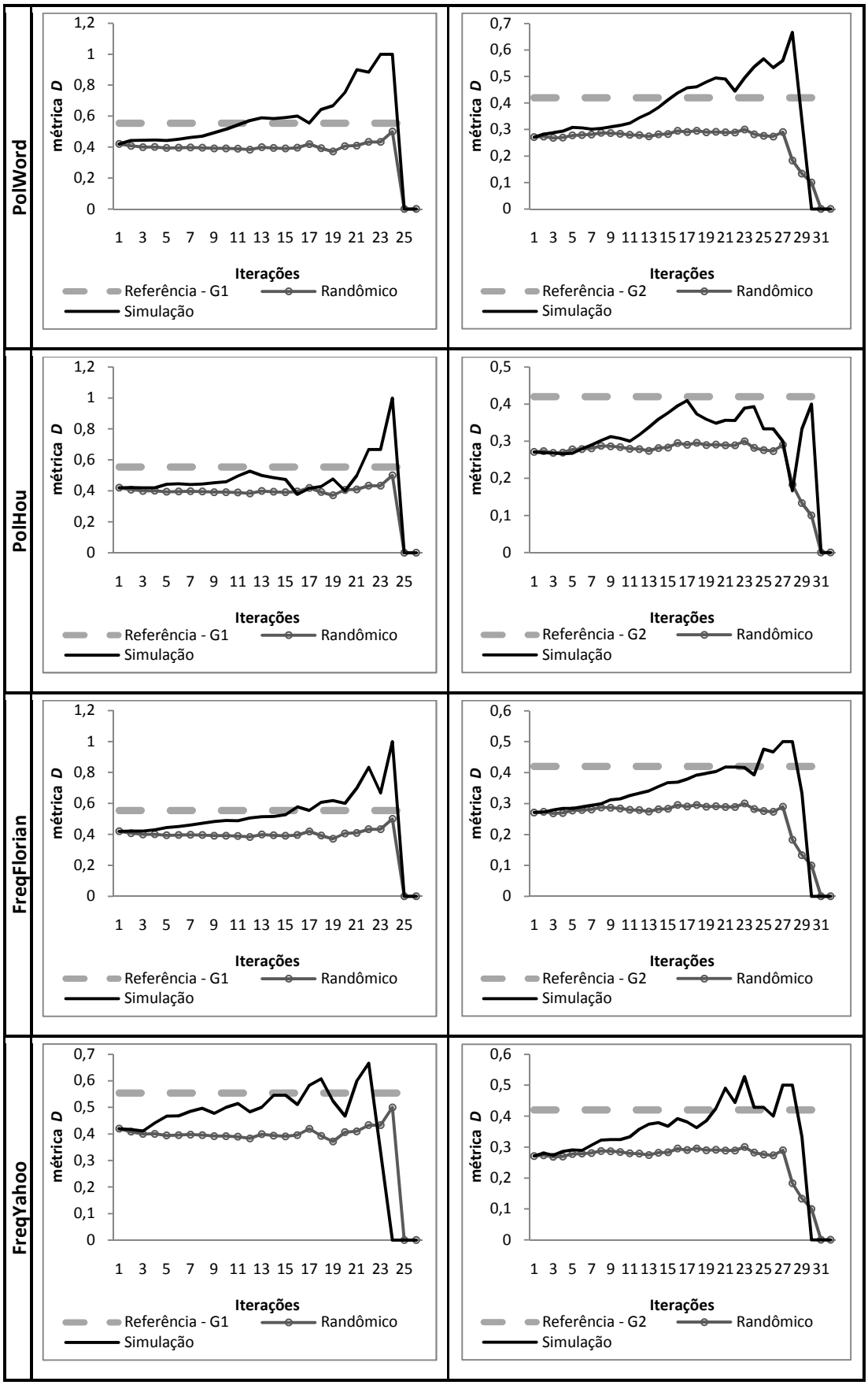


Figura 4.6: Evolução da métrica ‘conectividade média’ (<*k*>) nas duas simulações.

A evolução dos fatores lingüísticos em relação à densidade (*D*) foi ótima, com uma boa aproximação em direção à linha de referência no início, um período de instabilidade e uma piora no final (figura 4.7), para cima ou para baixo, na maior parte dos gráficos. Mais uma vez, é possível perceber um início próximo do randômico, mas uma maior aproximação da referência na primeira metade do gráfico. Esse início quase randômico deve-se à construção gradual do núcleo fortemente conectado, como abordado na análise da métrica <*k*>.




PolWord
PolHou
FreqFlorian
FreqYahoo
métrica D
Iterações
Referência - G1
Referência - G2
Randômico
Simulação

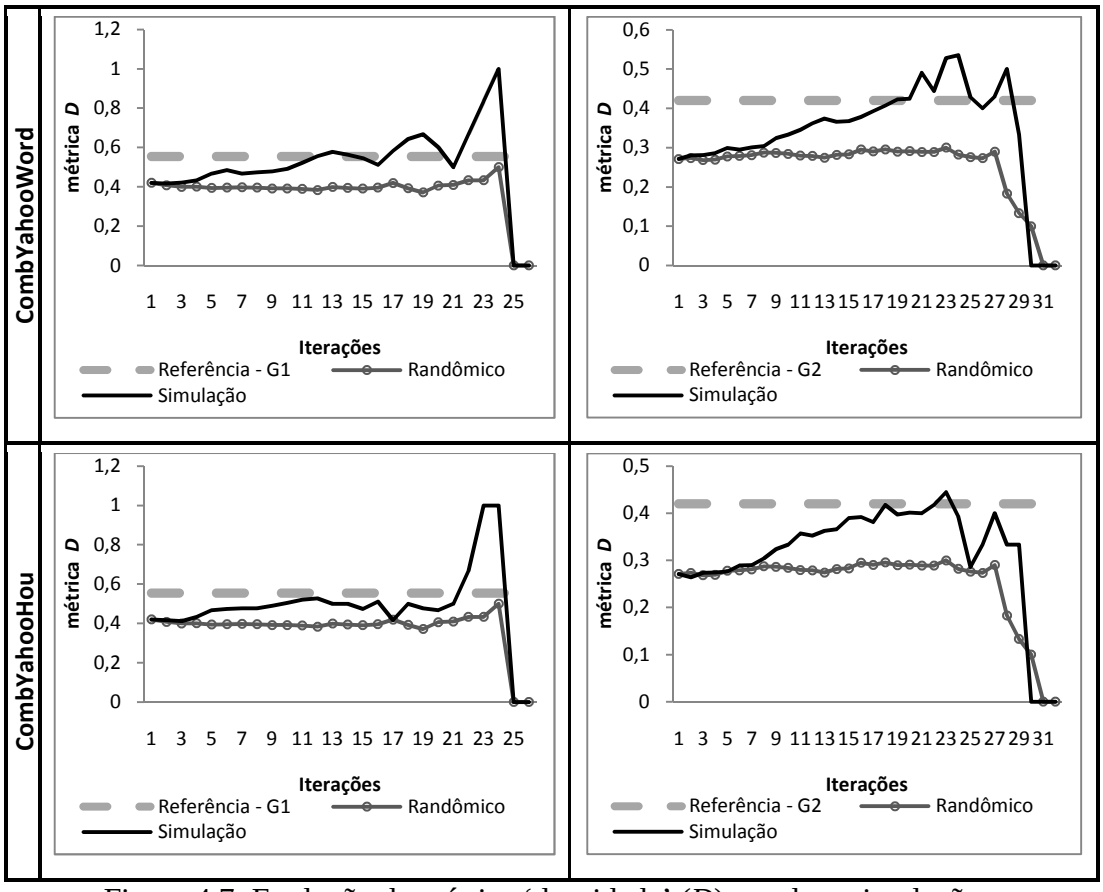


Figura 4.7: Evolução da métrica ‘densidade’ (*D*) nas duas simulações.

De um modo geral, as simulações apresentaram um comportamento de acordo com as hipóteses formuladas no capítulo 3. Em ambas, com todos os parâmetros lingüísticos testados, as maiores aproximações em relação à linha de referência, e o maior distanciamento em relação à linha randômica ocorreram na primeira metade da simulação. A maior exceção deu-se em relação ao caminho mínimo médio (*L*), por motivos explicados anteriormente e que serão abordados novamente na final deste capítulo. Quanto às combinações, o resultado indica uma simples mistura dos gráficos originais, sem que pareça ter havido uma otimização de fato.

4.2.2 Resultados das Métricas de Teoria dos Conjuntos

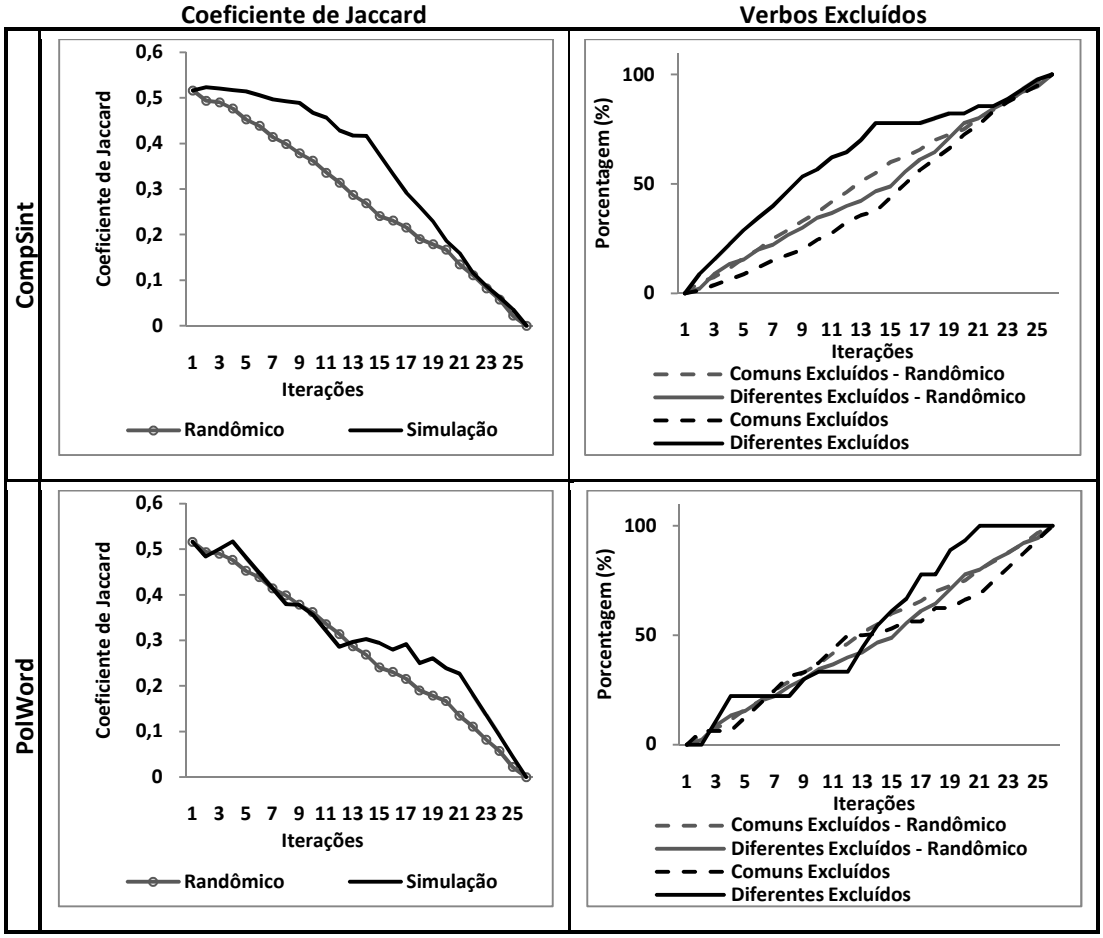
As métricas de teoria dos conjuntos serão apresentadas em dois gráficos: uma para o coeficiente de Jaccard e outro consolidando os verbos comuns aos dois grafos (componente “*x*” da fórmula) e os presentes apenas no grafo sendo modificado (componente “*z*”), como mencionado na seção 2.3.2. Os verbos denominados ‘diferentes’ são aqueles que se objetiva eliminar no início, uma vez que são eles que diferenciam um conjunto do outro, enquanto os verbos ‘comuns’, são os que esperamos eliminar no final, são os presentes em ambos os conjuntos.

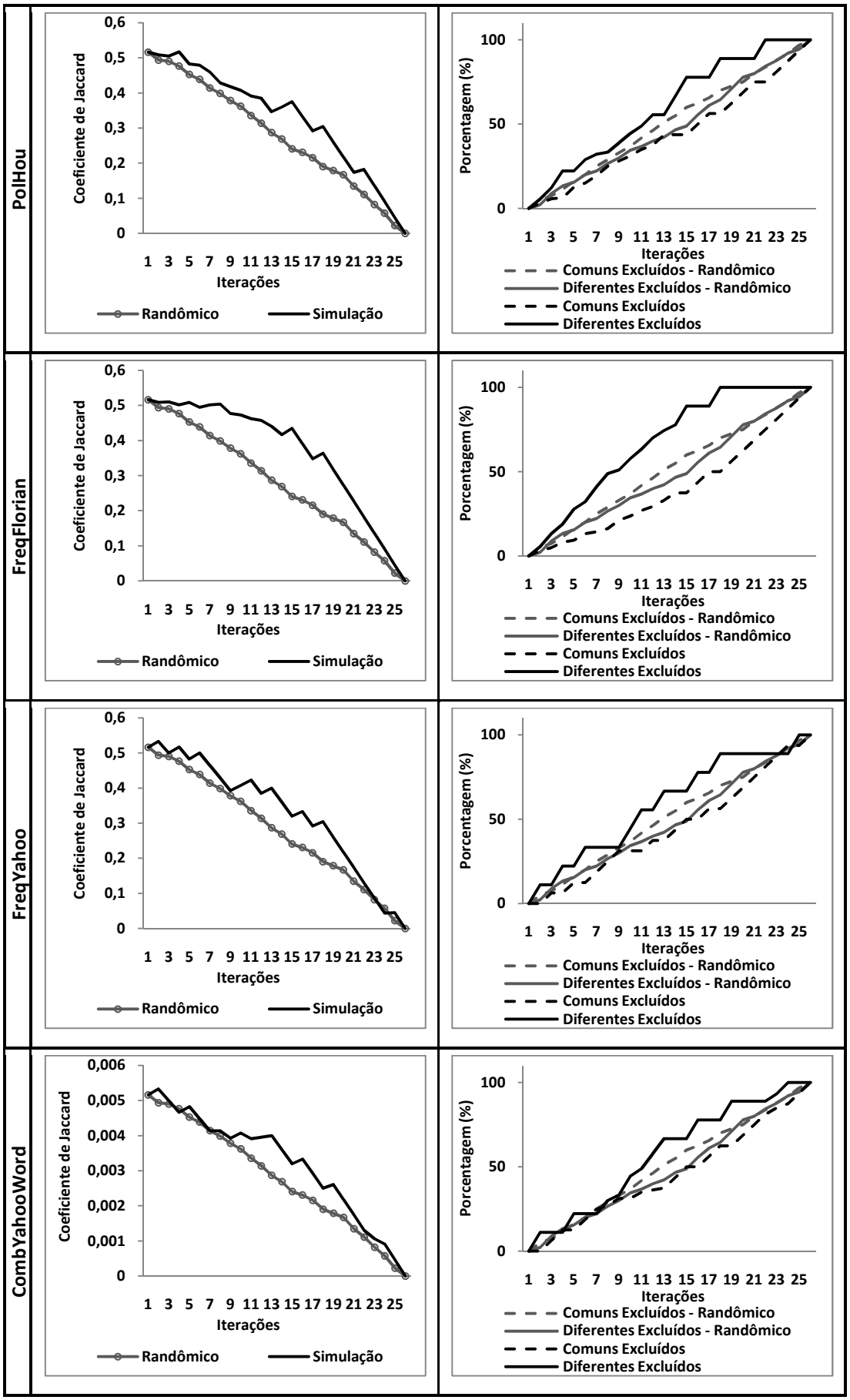
Uma vez que cada gráfico mostra aspectos da mesma informação, optou-se por uma apresentação diferenciada: uma simulação após a outra (primeiro alterando G2, depois alterando G3), com ambos os gráficos referentes ao mesmo escore lado a lado. Em cada gráfico, para cada linha, são apresentadas as equivalentes da eliminação randômica (média de 10 simulações, como explicado na seção anterior).

Nos gráficos apresentados a seguir, não existe uma linha de referência, uma vez que estamos executando uma comparação direta entre conjuntos de dados, não buscando um valor determinado. O objetivo, portanto, é um aumento direto na similaridade. No contexto do coeficiente de Jaccard, isso significa um aumento no valor, ou uma estagnação. Uma inclinação similar à randômica significa que uma quantidade proporcionalmente igual (ao randômico) de verbos ‘comuns’ e ‘diferentes’ está sendo realizada, enquanto inclinações mais horizontais ou aclives indicam uma quantidade maior (do que o randômico) de verbos ‘diferentes’ sendo eliminados. As linhas iniciam com o valor de similaridade dos dois grafos originais e terminam em 0.

Os gráficos dos verbos excluídos serão apresentados em percentual de eliminação, visto que a quantidade de verbos ‘diferentes’ é muito menor do que a de verbos ‘comuns’. Uma vez que declives não são possíveis, uma inclinação indica uma eliminação maior de verbos correspondentes à linha, enquanto uma estagnação (evolução horizontal) indica que eliminações de verbos daquele tipo não estão ocorrendo. Todas as linhas iniciam em 0% e terminam em 100%.

Nas duas simulações (figuras 4.8 e 4.9), CompSint e FreqFlorian apresentaram a mesma tendência inicial. Esse resultado já era esperado, uma vez que as listas dos dois escores possuem o exato mesmo conjunto inicial de verbos: aqueles que não foram mencionados no corpus ficaram empatados na primeira posição (13 e 20 verbos na primeira e segunda simulações respectivamente). Quanto a CompSint, o resultado da segunda simulação ficou muito próximo do randômico, enquanto o da primeira apresentou um resultado surpreendente positivo: dos 9 verbos 'diferentes', 7 foram englobados pelos verbos empatados no início.




PolHou
FreqFlorian
FreqYahoo
CombYahooWord
Coeficiente de Jaccard
Iterações
Randômico
Simulação
Porcentagem (%)
Comuns Excluídos - Randômico
Diferentes Excluídos - Randômico
Comuns Excluídos
Diferentes Excluídos

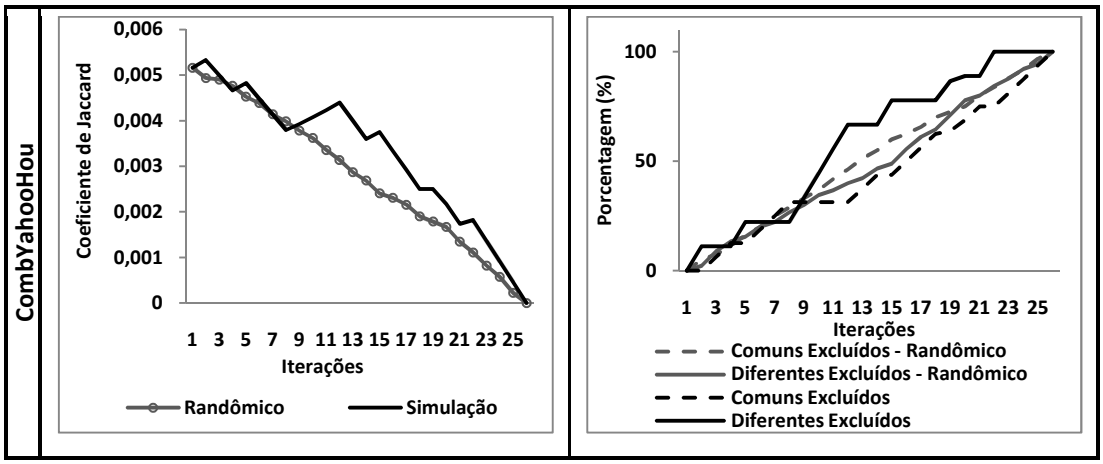
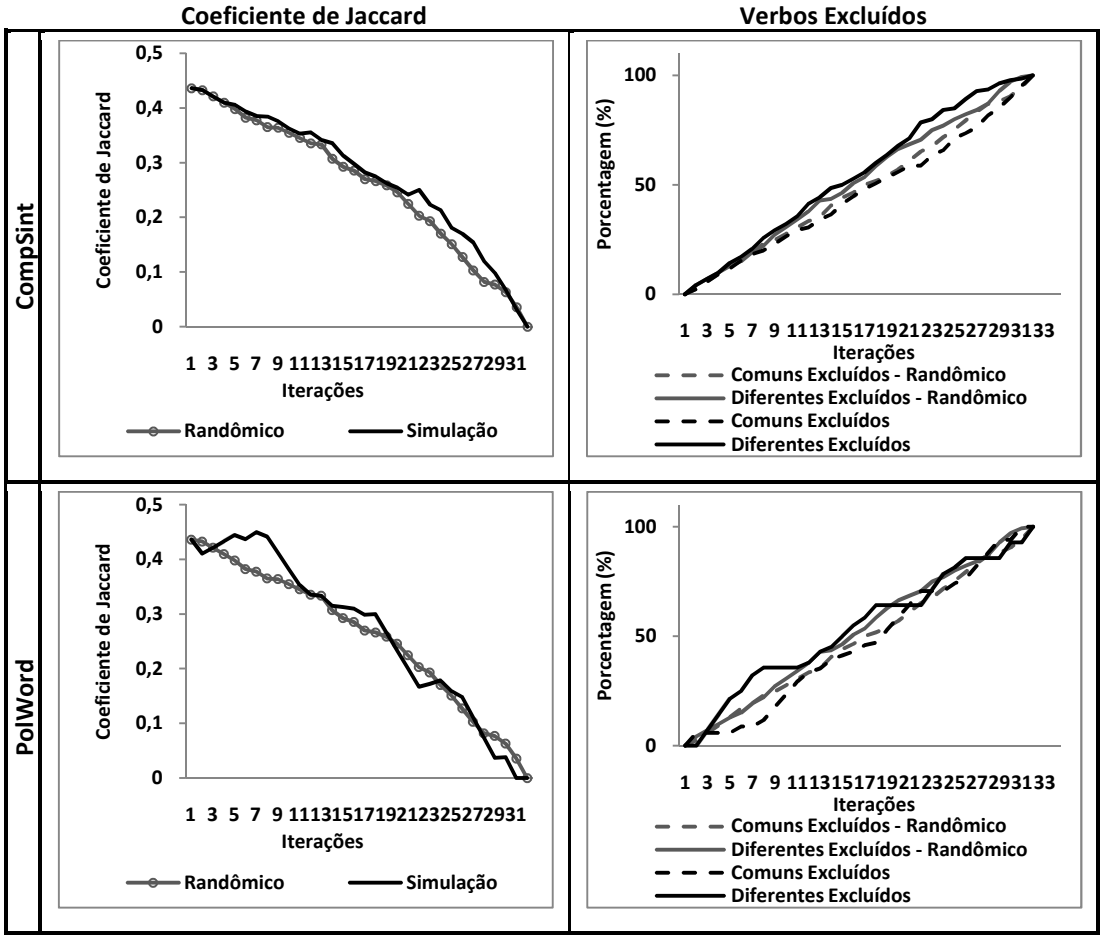
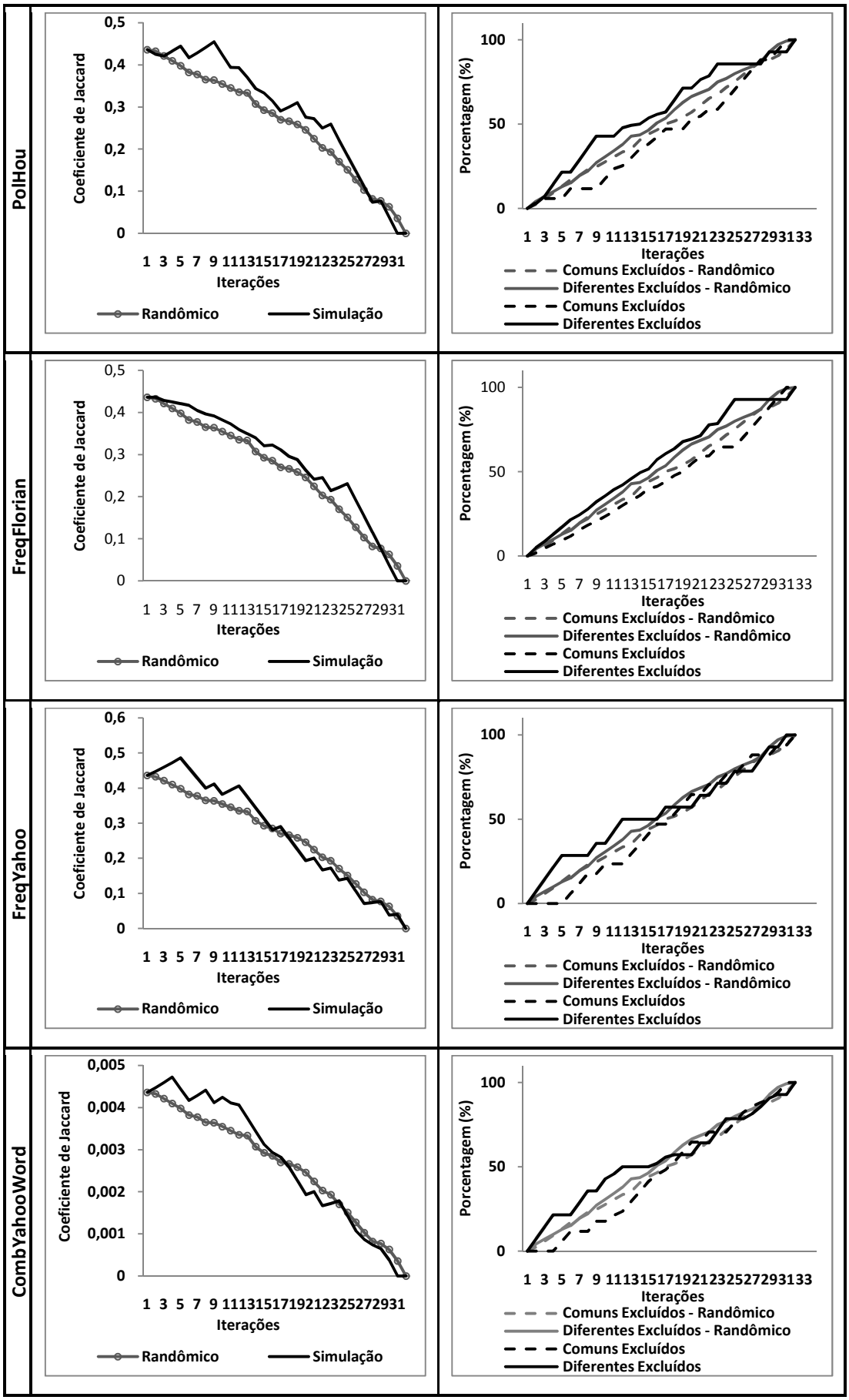


Figura 4.8: Evolução das métricas de teoria dos conjuntos na simulação alterando G2.




PolHou
FreqFlorian
FreqYahoo
CombYahooWord
Coeficiente de Jaccard
Iterações
Randômico
Simulação
Porcentagem (%)
Comuns Excluídos - Randômico
Diferentes Excluídos - Randômico
Comuns Excluídos
Diferentes Excluídos

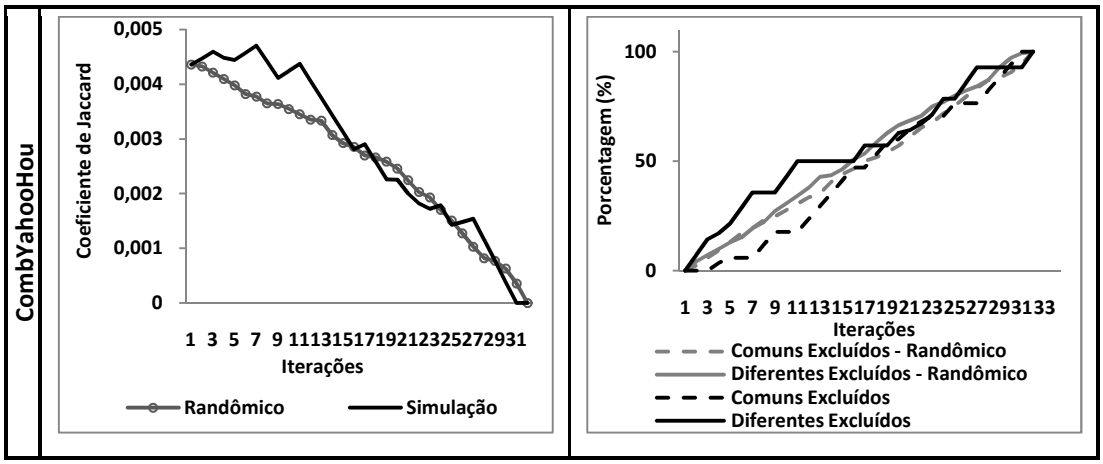


Figura 4.9: Evolução das métricas de teoria dos conjuntos na simulação alterando G3.

Quanto à polissemia, na primeira simulação, os resultados são positivos no sentido de apresentarem mais verbos excluídos 'diferentes' do que 'comuns' no início. Além disso, a linha dos excluídos permanece acima da dos 'comuns' durante a maior parte da simulação. Entretanto, no total, a maior quantidade de verbos 'diferentes' foi excluída mais para perto do fim, o que contraria as hipóteses sendo testadas. Por outro lado, os aclives acontecem em regiões centrais da simulação, não invalidado completamente, assim, a ordenação utilizada. Na segunda simulação, ambos os fatores lingüísticos apresentaram uma quantidade maior de verbos 'diferentes' excluídos no início, constituindo um resultado perfeitamente dentro das hipóteses formuladas.

A freqüência apresentou bons resultados, confirmando, em geral, o desenvolvimento esperado. O pior resultado foi o do escore FreqFlorian na segunda simulação, bastante próximo do randômico. Ainda assim, a quantidade de verbos 'diferentes' excluídos ficou acima das linhas da eliminação randômica e dos 'comuns' excluídos, durante a maior parte da simulação. A eliminação pelo escore FreqYahoo na primeira simulação, apresentou um comportamento ambíguo, com uma descida pouco antes do meio, seguida por um aclive acentuado. Uma vez que o maior declive acontece apenas no final, interpretamos este primeiro como ruído nos dados. Por outro lado, os resultados de FreqFlorian na primeira simulação e FreqYahoo na segunda foram particularmente positivos. Os gráficos ilustram um distanciamento significativo da linha randômica (i.e., a maior parte dos verbos 'diferentes' é eliminada no início) em ambos os casos, com um decréscimo apenas perto do meio ou final.

Quanto às combinações, o resultado parece ter sido novamente uma simples combinação de gráficos em vez de uma otimização. Particularmente em relação ao formato do gráfico, parece haver uma predominância da eliminação pelo fator FreqYahoo, com o gráfico levemente deformado de acordo com as respectivas polissemias. Na primeira simulação, percebe-se uma sensível piora em relação aos parâmetros misturados, em especial em relação ao escore PolHou. Na segunda simulação, é possível perceber uma melhora sutil, com mais verbos 'diferentes' excluídos antes da primeira metade da simulação.

4.2.3 Discussão

De um modo geral, os resultados foram ao encontro das hipóteses formuladas, tanto durante as análises preliminares, sobre os grafos dos grupos, quanto nas simulações de involução. Quanto aos resultados inesperados, alguns deles parecem diretamente

relacionados aos dados utilizados, em especial no que se refere a decisões tomadas no processo de coleta dos escores e de construção dos grafos.

Quanto ao problema da coleta, podemos citar a escolha do corpus Florianópolis. Como mencionado anteriormente, a complexidade sintática acabou subutilizada, uma vez que os primeiros verbos a serem eliminados foram aqueles onde não foi possível uma avaliação (por não terem sido mencionados no corpus). Em outras palavras, utilizou-se, na prática, a freqüência, ainda que assumindo uma provável complexidade alta.

Quanto ao problema da construção, podemos citar os resultados inesperados na avaliação dos escores médios por grafo, por não haver uma ponderação em relação ao quão mencionado foi determinado verbo. Na tabela 4.3, o impacto foi um aumento inesperado em PolWord e FreqYahoo em G2, contradizendo a tabela 4.4, onde a ponderação foi feita (ao considerar cada resposta na determinação da média do fator no grupo). Entretanto, os efeitos da ausência dessa valorização nas simulações são incertos. Se o impacto da eliminação de um verbo fosse proporcional à quantidade de vezes em que foi mencionado, é provável que a medição fosse mais precisa.

Outra dificuldade resultante da construção dos grafos é a baixa precisão das simulações por conta da pequena quantidade de vértices. O menor grafo modificado foi G2, com 25 vértices. Se considerarmos que estamos focando especificamente na primeira metade da simulação, sobram apenas 12 ou 13 iterações a serem consideradas. Adicionalmente, se desconsiderarmos as iterações nos extremos (início e fim) do gráfico (um procedimento comum na estatística, que visa o aumento da confiabilidade), o número reduz-se ainda mais. Dessa forma, é difícil traçar conclusões a respeito de simulações com efeitos importantes no meio, uma vez que uma pequena modificação na evolução poderia determinar a mudança da seção onde ocorre (primeira ou segunda metade). Como exemplos podemos citar as métricas de teoria dos conjuntos para PolWord e PolHou na primeira simulação, onde o maior incremento acontece no meio, e FreqYahoo, na primeira simulação, juntamente com PolHou, na segunda, onde existe um vale nessa posição central.

Por fim, existem dificuldades relacionadas ao método de interpretação em si, ao menos no que tange a teoria dos grafos. A dinâmica formulada originalmente consistia em eliminar os vértices iterativamente e verificar se as métricas coletadas aproximavam-se da linha de referência. Entretanto, por características do próprio grafo, por vezes a interpretação acaba não sendo não-trivial, como a evolução quase randômica de L na segunda simulação (em todos os escores). Além das características do grafo, o elevado particionamento na eliminação randômica foi preponderante não apenas para o comportamento inesperado de L, mas também de *diam*. Os caminhos mínimos foram computados considerando apenas os vértices alcançáveis, portanto os subgrafos pequenos acabam tendendo a diminuir L e *diam* de uma forma global (os alcançáveis estão mais perto uns dos outros). Uma vez que L e *diam* são métricas de difícil alteração em G3 (como explicado anteriormente), o particionamento do grafo na eliminação randômica resulta em uma diminuição artificial de destas métricas. A figura 4.10 ilustra o problema, mostrando exemplos de grafos gerados por meio de eliminação randômica em iterações próximas da metade do processo de simulação.

Figura 4.10: Exemplos de grafos da eliminação randômica: alterando G2, iteração 12 (três subgrafos) e alterando G3, iteração 15 (dois subgrafos) respectivamente.

O aparecimento de partições favorece a diminuição de *L* e *diam* na eliminação randômica (ao gerar subgrafos, diminuindo os caminhos entre os nodos alcançáveis), o que é considerado um bom resultado na segunda eliminação (visto que a linha de referência encontrava-se abaixo do ponto de partida da simulação). Um conjunto de dados maior, em especial com uma quantidade maior de arestas, que dificultasse o aparecimento de partições, poderia produzir resultados mais interessantes.

5 CONCLUSÕES E TRABALHOS FUTUROS

A presente pesquisa abordou a influência de três fatores lingüísticos na evolução lexical verbal: freqüência de observação, polissemia e complexidade sintática. A escolha dos fatores deu-se por sua comprovada importância cognitiva e por eventuais dificuldades associadas a outros fatores (em especial no que tange a polissemia, como visto na seção 2.2.2). Foram formuladas hipóteses de que verbos com maior freqüência, maior polissemia e menor complexidade sintática seriam adquiridos e utilizados mais cedo pelas crianças, e que esta tendência poderia ser verificada computacionalmente. Também foi assumido que a análise de combinações desses fatores poderia produzir resultados ainda melhores.

A investigação foi baseada no vocabulário de três grupos etários (TONIETTO, 2009; TONIETTO et al., 2008) transformados em grafos: G1, G2 e G3. A metodologia incluiu modificações iterativas dos grafos dos indivíduos mais velhos (G2 e G3) e comparações com o grafo imediatamente mais novo (G1 e G2 respectivamente). O processo de modificação proposto foi chamado de 'involução', um procedimento similar ao *network growth* (ALBERT & BARABÁSI, 2002; STEYVERS & TENENBAUM, 2005), porém inverso: os nodos não são adicionados, mas eliminados do grafo, numa ordem que depende dos fatores lingüísticos testados. As análises incluíram métricas da teoria dos grafos (*S*, *C/s*, *L*, *diam*, *<k>* e *D*) e da teoria dos conjuntos (coeficiente de Jaccard e suas componentes).

A análise dos grafos de cada conjunto sugeriu a comprovação das hipóteses, indicando diminuição da freqüência média e da polissemia média, bem como aumento na complexidade sintática média conforme aumenta a idade média dos grupos. Isso pôde ser comprovado tanto em análises preliminares das métricas (tabelas 4.1 e 4.2), quanto nas médias dos escores lingüísticos por grupo (tabelas 4.3 e 4.4) e também em uma ANOVA (anexo A). As exceções encontradas na análise dos escores lingüísticos do grafo em si (tabela 4.3) foram creditadas a uma ausência de ponderação: cada verbo do grafo contribuiu igualmente para o resultado. Quando a quantidade de menções foi levada em consideração (tabela 4.4), as hipóteses foram comprovadas completamente.

A análise das simulações também indicou, em geral, uma confirmação das hipóteses. Algumas exceções, discutidas na seção 4.2.3, devem-se, a problemas com os dados, tanto na construção dos grafos quanto na coleta dos fatores lingüísticos. Adicionalmente, foi possível perceber que a análise das simulações nem sempre é simples. A proposta inicial da presente pesquisa era verificar se a manipulação de grafos, segundo os fatores lingüísticos, resultaria em um aumento de similaridade (relativo a conteúdo e estrutura) do grafo alterado em relação ao objetivo. Entretanto, uma estabilidade em *diam*, por exemplo, é considerada positiva, uma vez que aponta para a manutenção estrutural, ainda que não necessariamente levando a um aumento de similaridade. Também ficou evidente a importância da adaptação de algumas métricas à

presença de subgrafos, destacando-se a evolução insatisfatória de *L* na segunda simulação, e os bons resultados de *C/s*, uma versão adaptada do coeficiente de clusterização original (*C*).

Quanto às composições de fatores lingüísticos, os resultados apontam para um desempenho mediano. Era esperado que, ao combinar a influência da polissemia e da freqüência, um resultado melhor seria verificado. Entretanto, parece ter havido uma simples mistura dos gráficos, por vezes pior do que alguma das partes componentes.

Em suma, tantos os resultados preliminares quanto as simulações sugerem uma confirmação da influência dos fatores lingüísticos conforme assumido no início da desta pesquisa. Estes fatores parecem não apenas influenciar fortemente o conteúdo do vocabulário dos diferentes grupos, como também contribuir para a estruturação do léxico-mental. Adicionalmente, a análise foi útil para evidenciar algumas propriedades interessantes, como o aumento do vocabulário com a idade, bem como o aumento da especialização. Também demonstrou que os grafos das crianças (G1 e G2) são mais similares entre si do que em relação ao dos adultos (G3), tanto no que se refere à estrutura quanto ao conteúdo. Em vista disso, concluímos que tanto o procedimento de involução proposto, quanto a análise realizada são apropriados para estudos lingüísticos referentes à evolução do vocabulário verbal.

5.1 Trabalhos Futuros

Alguns dos resultados da presente pesquisa foram parcialmente comprometidos por conta de escolhas equivocadas e pela própria característica de alguns dados (como visto na seção 4.2.3). Como trabalho futuro, pretendemos testar novamente a complexidade sintática, dessa vez utilizando outro corpus, maior, que permita a verificação plena da influência do fator lingüístico. Da mesma forma, testes de um novo escore de freqüência sobre um corpus mais amplo pode ser proveitoso.

Adicionalmente, planejamos ampliar o tamanho dos grafos, complementando os dados existentes com uma nova aplicação do experimento de nomeação. Também objetivamos reestruturá-los, de modo que seja possível expressar a importância de cada verbo dentro do vocabulário (proporcional à quantidade de vezes em que foi mencionado).

Por fim, pretendemos aplicar a mesma análise sobre outros fatores lingüísticos, como concretude. Pretendemos também utilizar a metodologia para analisar dissolução léxica no contexto de patologias como o mal de Alzheimer.

REFERÊNCIAS

ALBERT, R., BARABÁSI, A.-L. Statistical mechanics of complex networks. **Reviews of Modern Physics**, [S.l.], v.74, n.1, p. 47-97, Jan. 2002.

ANTIQUEIRA, L.; NUNES, M.G.V.; OLIVEIRA JR, O. N; COSTA, L da F. Strong correlations between text quality and complex networks features. **Physica A: Statistical Mechanics and its Applications**. [S.l.], v.373, p. 811-820, Jan. 2007.

BARDE, L. H. F.; SCHWARTZ, M. F., BORONAT, C. B. Semantic weight and verb retrieval in aphasia. **Brain and Language**. [S.l.], v.97, n.3, p. 266-278, Jun. 2006

BATAGELJ, V.; MRVAR, A. **Pajek – Program for Large Network Analysis**. Disponível em <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>. Acesso em jan. 2009.

BLACK, M.; CHIAT, S. Noun–verb dissociations: a multi-faceted phenomenon. **Journal of Neurolinguistics**, [S.l.], v.16, n.2-3, p. 231-250, Mar-May. 2003.

BREEDIN, S.; SAFFRAN, E. M.; SCHWARTZ, M. F. Semantic Factors in Verb Retrieval: An Effect of Complexity. **Brain and Language**, New York, v.63, n.1, p. 1-31, Jun. 1998.

CORONGES, K. A.; STACY, A. W.;VALENTE, T. W. Structural Comparison of Cognitive Associative Networks in Two Populations. **Journal of Applied Social Psychology**, [S.l.], v. 37, n.9, p. 2097-2129. 2007.

DAVIDOFF, J.; MASTERSON, J. The development of picture naming: Differences between verbs and nouns. **Journal of Neurolinguistics**, [S.l.], v.9, n.2, p. 69-83, Apr. 1996.

DE DEYNE, S; STORMS, G. Word associations: Network and semantic properties. **Behavior Research Methods**, [S. l.], v. 40, n. 1, p. 213-231. 2008.

DIAS-DA-SILVA, B.C. et al. Construção de um thesaurus eletrônico para o português do Brasil. In: PROCESSAMENTO COMPUTACIONAL DO PORTUGUÊS ESCRITO E FALADO, PROPOR, 4., 2000, [S. l.]. **Anais…** [S. l.:s.n.], 2000. p. 1-10.

DIAS-DA-SILVA, B.C.; MORAES, H.R. A construção de um thesaurus eletrônico para o português do Brasil. **ALFA: Revista de Lingüística**, [S. l.], v. 47, n. 2, p. 101-115, jul. 2003.

DOROW, B.; WIDDOWS, D. Discovering corpus-specific word senses. In: EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, EACL, 2003. **Proceedings**... Morristown, NJ, USA: Association for Computational Linguistics, 2003. p. 79-82.

ELLIS, A. W.; MORRISON, C. N. Real Age-of-Acquisition Effects in Lexical Retrieval. **Journal of Experimental Psychology: Learning, Memory, and Cognition**, [S.l.], v.24, n.2, p. 515-523. 1998.

ELLIS, A.; RALPH, M. A. L. Age of Acquisition Effects in Adult Lexical Processing Reflect Loss of Plasticity in Maturing Systems: Insights From Connectionist Networks. **Journal of Experimental Psychology: Learning, Memory, and Cognition**, [S.l.], v.26, n.5, p. 1103-1123. 2000.

EVANS, V., GREEN, M. **Cognitive Linguistis: An Introduction**. Edinburgh: Edinburgh University Press, 2006.

FAZLY, A.; ALISHAHI, A.; STEVENSON, S. A Probabilistic Incremental Model of Word Learning in the Presence of Referential Uncertainty. In: CONFERENCE OF THE COGNITIVE SCIENCE SOCIETY, CogSci, 30., 2008. **Proceedings**... Washington, DC: [s.n.], 2008.

FERRER i CANCHO, R.; SOLÉ, R. V.; KÖHLER, R. Patterns in syntactic dependency networks. **Phys. Rev. E**. [S.l.], v.69, n.5, May. 2004.

GAUME, B.; VENANT, F.; VICTORRI, B. Hierarchy in Lexical Organization of Natural Languages. In: Pumain, D. (Ed.). **Hierarchy in Natural and Social Sciences**. [S.l.]: Springer Netherlands, 2006. p. 121-142.

GERMANN, D. C.; VILLAVICENCIO, A.; SIQUEIRA, M. An Investigation on Polysemy and Lexical Organization of Verbs. In: WORKSHOP ON COMPUTACIONAL LINGUISTICS, NAACL HLT, 1., 2010. **Proceedings...** Los Angeles, California, USA:[s.n.], 2010-a.

GERMANN, D. C.; VILLAVICENCIO, A.; SIQUEIRA, M. An Investigation on the Influence of Frequency on the Lexical Organization of Verbs. In: TEXTGRAPHS WORKSHOP, ACL TEXTGRAPHS, 5., 2010. **Proceedings...** Uppsala, Sweden:[s.n.], 2010-b.

GOLDBERG, A. E. The Emergence of the Semantics of Argument Structure Constructions. In: MACWHINNEY, B. (Ed.), **Emergence of Language**. Mahwah, NJ: Lawrence Erlbaum Associates, 1999. p. 197-212.

GONG, T.; MINETT, J.W.; WANG, W. S-Y. Exploring social structure effect on language evolution based on a computational model. **Connection Science**, London, UK, v.20, n.2-3, p. 135–153. 2008.

GORMAN, J; CURRAN, J. R. The Topology of Synonymy and Homonymy Networks. In: WORKSHOP ON COGNITIVE ASPECTS OF COMPUTATIONAL LANGUAGE ACQUISITION, 2007. **Proceedings...** Prague, Czech Republic: Association for Computational Linguistics, 2007.

HORST, J. S.; MCMURRAY, B.; SAMUELSON, L. K. Online Processing is Essential for Learning: Understanding Fast Mapping and Word Learning in a Dynamic Connectionist Architec. In: CONFERENCE OF THE COGNITIVE SCIENCE SOCIETY, CogSci, 28., 2006. **Proceedings...** Nashville, TN: [s.n.], 2006.

HOUAISS, A. **Dicionário Eletrônico Houaiss da Língua Portuguesa**, versão 2.0a. [S.l.]: Editora Objetiva. 2007.

HOWES, D. H.; SOLOMON, R.L. Visual duration threshold as a function of word-probability. **Journal of Experimental Psychology**, [S.l.], v.41, n.6, p. 401-410, Jun. 1951.

JACCARD, P. Distribution de la flore alpine dans le Bassin des Drouces et dans quelques re-gions voisines. **Bulletin de la Société Vaudoise des Sciences Naturelles**, [S.l.], v. 37, n.140, p. 241–272. 1901.

JACKENDOFF, R. **Patterns in the Mind: Language and human nature**. [S. l.]: BasicBooks. 1994.

KIM, M.; THOMPSON, C. K. Verb deficits in Alzheimer's disease and agrammatism: Implications for lexical organization. **Brain and Language**. [S.l.], v.88, n.1, p. 1-20, Jan. 2004.

KIRAN, S.; THOMPSON, C. K. The Role of Semantic Complexity in Treatment of Naming Deficits: Training Semantic Categories in Fluent Aphasia by Controlling Exemplar Typicality. **Journal of Speech, Language, and Hearing Research**, [S.l.], v.46, n.3, p. 608-622, Jun. 2003.

LAAKSO, L.; SMITH, L. B. Pronouns and verbs in adult speech to children: A corpus analysis. **Journal of Child Language**, [S. l.], v.34, n.4, p. 725-763, Oct. 2007.

LAKOFF, G. **Women, Fire, and Dangerous Things**. Chicago: University of Chicago Press. 1987.

LAUDANNA, A.; GAZZELLINI, S.; DE MARTINO, M. Representation of grammatical properties of Italian verbs in the mental lexicon. **Brain and Language**. [S.l.], v.90, n.1-3, p. 95-105, Jul-Sep. 2004.

LI, P.; FARKAS, I.; MACWHINNEY, B. Early lexical development in a self-organizing neural network. **Neural Networks**, Oxford, UK, v.17, n.8-9, p.1345-1362. 2004.

MACWHINNEY, B. **The CHILDES project: Tools for analyzing talk.** 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

MANDLER, Jean M.. Conceptual Categorization. In: RAKISON, D. H.; OAKS, L. M. (Comp.). **Early Category and Concept Development: Making Sense of the Blooming, Buzzing Confusion**. Nova York: Oxford University Press, 2003. p. 103-131. Disponível em: <http://site.ebrary.com/lib/ufrgs/Doc?id=10084848&ppg=6>. Acesso em: 18 jul. 2008.

MARCUS, G. F., VIJAYAN, S., BANDI RAO, S., VISHTON, P. M.. Rule learning by seven-month-old-infants**. Science**, [S.l], n. 283. p. 77-80. 1999.

MASUCCI, A. P; RODGERS, G. J. Network properties of written human language. **Physical Review E**, [S.l.], v. 74, n.2, Aug. 2006.

MAZIERO, E.G. et al. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In: WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, TIL, 6., 2008, Vila Velha. **Anais…** Vila Velha: 2008. p. 390-392.

MEHLER, Jacques. Language Related Dispositions in Early Infancy. In: MEHLER, J.; FOX, R. (Comp.). **Neonate Cognition: Beyond the Blooming Buzzing Confusion**. Nova Jersey: Lawrence Erlbaum Associates Publishers, 1985. p. 7-26.

MEHLER, A. Evolving Lexical Networks. A Simulation Model of Terminological Alignment. In: EUROPEAN SUMMER SCHOOL IN LOGIC, LANGUAGE AND INFORMATION, ESSLLI, 19., 2007. **Workshop Proceedings...** Trinity College, Dublin:[s.n.], 2007.

MELTZOFF, A. N., MOORE, M. K. Cognitive Foundations and Social Functions of Imitation and Intermodal Representation in Infancy. In: MEHLER, J.; FOX, R.

(Comp.). **Neonate Cognition: Beyond the Blooming Buzzing Confusion**. Nova Jersey: Lawrence Erlbaum Associates Publishers, 1985. p. 139-156.

MERVIS, C.. Early Lexical Development: The Contributions of Mother and Child. In: SOPHIN, C. (Ed.) **Origins of Cognitive Skills**. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1984. p. 339-370.

MORRISON, C. M.; ELLIS, A. W. Roles of Word Frequency and Age of Acquisition in Word Naming and Lexical Decision. **Journal of Experimental Psychology: Learning, Memory, and Cognition**, [S.l], v.21, n.1, p. 116-133. 1995.

MOTTER, A. E.; DE MOURA, A. P S.; LAI, Y.-C.; DASGUPTA, P. Topology of the conceptual network of language, **Physical Review E**, [S.l.], v.65. 2002.

NAVIGLI, R.; LAPATA, M. Graph Connectivity Measures for Unsupervised Word Sense Disambiguation. In: International Joint Conference on Artificial Intelligence, 20., 2007. **Proceedings...** Hyderabad, India: [s.n.], 2007. p. 1683-1688.

NELSON, D. L.; MCKINNEY, V. M.; GEE, N. R.; JANCZURA, G. A. Interpreting the influence of implicitly activated memories on recall and recognition. **Psychological Review**, [S.l.], v.105, n.2, p. 299-324. 1998.

PARISIEN, C.; STEVENSON, S. Modelling the acquisition of verb polysemy in children. In: Workshop on Distributional Semantics beyond Concrete Concepts , DiSCo, 2009. **Workshop Proceedings...** [S.l.:s.n.]. 2009. p. 17-22.

PLAUT, D. Relearning after Damage in Connectionist Networks: Toward a Theory of Rehabilitation. **Brain and Language**. New York, v.52, n.1, p. 25-82, Jan. 1996.

REGIER, T. The emergence of words: Attentional learning in form and meaning. **Cognitive Science: A Multidisciplinary Journal**, London, UK, v.29, n.6, p. 819–865. 2005.

SCHMID, H., UNGERER, F. **An Introduction to Cognitive Linguistics.** 2nd ed. Harlow: Pearson Education Limited. 2006.

SCHOENEMANN, P. T. Syntax as an Emergent Characteristic of the Evolution of Semantic Complexity. **Minds and Machines**, Hingham, MA, USA: Kluwer Academic Publishers, v.9, n.3, p.309-346, Aug. 1999.

SCLIAR-CABRAL, L. **Corpus Florianópolis**. 1993. Disponível em: <http://childes.psy.cmu.edu/data/Romance/Portuguese/Florianopolis.zip>. Acesso em: jan. 10 2009.

SIGMAN, M.; CECCHI. G. A. Global organization of the WordNet lexicon. **Proceedings of the National Academy of Sciences of the United States of America**, [S.l.], v.99, n.3, Feb. 2002.

SINHA, R.; MIHALCEA, R. Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity. In: IEEE International Conference on Semantic Computing, ICSC, 2007. **Proceedings...** Irvine, CA: [s.n.], 2007.

SISKIND, J. M. A computational study of cross-situational techniques for learning word-to-meaning mappings. **Cognition**, [S.l.], v.61, n.1-2, p. 1-38, Oct-Nov. 1996.

SOARES, M. M.; CORSO, G.; LUCENA, L. S. The network of syllables in Portuguese. **Physica A: Statistical Mechanics and its Applications**, [S.l.], v. 355, n. 2-4, p. 678-684, Sep. 2005.

SOLOMON, R. L; POSTMAN, L. Frequency of usage as a determinant of recognition thresholds for words. **Journal of Experimental Psychology**, [S.l.], v.43, n. 3, p. 195-201. 1952.

SPELKE, E. Perception of Unity, Persistence, and Identity: Thoughts on Infants' Conceptions of Objects. In: MEHLER, J.; FOX, R. (Comp.). **Neonate Cognition: Beyond the Blooming Buzzing Confusion**. Nova Jersey: Lawrence Erlbaum Associates Publishers, 1985. p. 89-113.

STEYVERS, M.; TENENBAUM, J. B. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. **Cognitive Science: A Multidisciplinary Journal**, [S.l.], v.29, n.1, p. 41-78. 2005.

THOMPSON, C. K. Unaccusative verb production in agrammatic aphasia: the argument structure complexity hypothesis. **Journal of Neurolinguistics**, [S.l.], v.16, n.2-3, p. 151-167, Mar-May. 2003.

THOMPSON, C. K.; SHAPIRO, L. P.; KIRAN, S. The Role of Syntactic Complexity in Treatment of Sentence Deficits in Agrammatic Aphasia: The Complexity Account of Treatment Efficacy (CATE). **Journal of Speech, Language, and Hearing Research**, [S.l.], v.46, n.3, p. 591-607, Jun. 2003.

TOMASELLO, M. **Constructing a Language: A Usage-Based Theory of Language Acquisition**. Estados Unidos da América: Harvard University Press. 2003.

TONIETTO, L; VILLAVICENCIO, A.; SIQUEIRA, M.; PARENTE, M. A. P.; SPERB, T. A especificidade semântica como fator determinante na aquisição de verbos. **Psico**, [S.l.], v.39, n.3, p. 343-351, Jul-Set. 2008.

TONIETTO, L. **Desenvolvimento da convencionalidade e especificidade na aquisição de verbos: relações com complexidade sintática e categorização**. 2009. 199 f. Tese (Doutorado em Psicologia) - Instituto de Psicologia, UFRGS, Porto Alegre.

UZIEL-KARL, S. **A Multidimensional Perspective on the Acquisition of Verb Argument Structure**. 2001. 356 f. Tese (Doctor of Philosophy) – The Shirley & Leslie Porter School of Cultural Studies, Tel Aviv University, Tel Aviv, Israel.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of ’small-world’ networks. **Nature**, [S.l.], v.393, n.6684, p. 440-442, Apr. 1998.

WITTGENSTEIN, L.. **Philosophical Investigations**. New York: Macmillan, 1953.

XU, F.; TENENBAUM, J. B. Word learning as Bayesian inference. **Psychological Review**, USA, v.114, n.2, p. 245-272, Apr. 2007.

YOUNGER, Barbara A.. Parsing Objects into Categories: Infants' Perception and Use of Correlated Attributes. In: RAKISON, D. H.; OAKS, L. M. (Comp.). **Early Category and Concept Development: Making Sense of the Blooming, Buzzing Confusion**. Nova York: Oxford University Press, 2003. p. 77-102. Disponível em: <http://site.ebrary.com/lib/ufrgs/Doc?id=10084848&ppg=6>. Acesso em: 18 jul. 2008.

YU, C. The emergence of links between lexical acquisition and object categorization: A computational study. **Connection Science**, [S.l.], v.17, n.3-4, p. 381-397, Sep-Dec. 2005.

YU, C. Learning syntax–semantics mappings to bootstrap word learning. In: CONFERENCE OF THE COGNITIVE SCIENCE SOCIETY, CogSci, 28., 2006. **Proceedings...** Nashville, TN: [s.n.], 2006.

ANEXO A – RESULTADOS OBTIDOS PELO SOFTWARE SPSS

Descriptives

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
1	55	16,25319	3,862111	,520767	15,20912	17,29727
2	55	14,66492	2,509155	,338334	13,98660	15,34324
3	55	11,12587	2,349031	,316743	10,49084	11,76090
Total	165	14,01466	3,663638	,285214	13,45150	14,57783

Descriptives

	Minimum	Maximum
1	9,563	24,467
2	10,765	22,235
3	6,471	18,200
Total	6,471	24,467

Figura A.1: Estatísticas descritivas da polissemia extraída do tesauro WordNetBR.

Test of Homogeneity of Variances

Levene Statistic	df1	df2	Sig.
11,629	2	162	,000

Figura A.2: Teste de Levene da polissemia extraída do tesauro WordNetBR.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	757,843	2	378,922	42,528	,000
Within Groups	1443,404	162	8,910		
Total	2201,248	164			

Figura A.3: ANOVA da polissemia extraída do tesauro WordNetBR.

Multiple Comparisons

	(I) Grupo	(J) Grupo	Mean Difference (I-J)
Tukey HSD	1	2	1,588268*
		3	5,127321*
	2	1	-1,588268*
		3	3,539053*
	3	1	-5,127321*
		2	-3,539053*

*. The mean difference is significant at the 0.05 level

Figura A.4: Teste de Comparações Múltiplas de Tukey da polissemia extraída do tesauro WordNetBR.

Descriptives

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
1	55	26,93069	6,791920	,915822	25,09458	28,76680
2	55	23,01529	4,489276	,605334	21,80167	24,22891
3	55	17,82406	3,281874	,442528	16,93685	18,71128
Total	165	22,59001	6,274443	,488465	21,62552	23,55450

Descriptives

Minimum	Maximum
16,563	43,313
16,471	36,471
11,412	24,118
11,412	43,313

Figura A.5: Estatísticas descritivas da polissemia extraída do dicionário Houaiss.

Test of Homogeneity of Variances

Levene Statistic	df1	df2	Sig.
13,294	2	162	,000

Figura A.6: Teste de Levene da polissemia extraída do dicionário Houaiss.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2295,515	2	1147,758	44,686	,000
Within Groups	4160,941	162	25,685		
Total	6456,457	164			

Figura A.7: ANOVA da polissemia extraída do dicionário Houaiss.

Multiple Comparisons

	(I) Grupo	(J) Grupo	Mean Difference (I-J)	Std. Error	Sig.
Tukey HSD	1	2	3,915402*	,966433	,000
		3	9,106630*	,966433	,000
	2	1	-3,915402*	,966433	,000
		3	5,191228*	,966433	,000
	3	1	-9,106630*	,966433	,000
		2	-5,191228*	,966433	,000

Figura A.8: Teste de Comparações Múltiplas de Tukey da polissemia extraída do dicionário Houaiss.

Descriptives

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
1	55	4,74900	3,201780	,431728	3,88344	5,61457
2	55	6,60877	1,852037	,249729	6,10810	7,10945
3	55	10,70105	1,961699	,264515	10,17073	11,23137
Total	165	7,35294	3,462807	,269579	6,82065	7,88524

Descriptives

	Minimum	Maximum
1	,063	12,400
2	1,824	10,941
3	6,353	15,412
Total	,063	15,412

Figura A.9: Estatísticas descritivas da complexidade sintática.

Test of Homogeneity of Variances

Levene Statistic	df1	df2	Sig.
10,602	2	162	,000

Figura A.10: Teste de Levene da complexidade sintática.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1019,926	2	509,963	87,274	,000
Within Groups	946,604	162	5,843		
Total	1966,530	164			

Figura A.11: ANOVA da complexidade sintática.

Multiple Comparisons

	(I) Grupo	(J) Grupo	Mean Difference (I-J)	Std. Error	Sig.
Tukey HSD	1	2	-1,859771*	,460957	,000
		3	-5,952047*	,460957	,000
	2	1	1,859771*	,460957	,000
		3	-4,092276*	,460957	,000
	3	1	5,952047*	,460957	,000
		2	4,092276*	,460957	,000

*. The mean difference is significant at the 0.05 level.

Figura A.12: Teste de Comparações Múltiplas de Tukey da complexidade sintática.

Descriptives

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
1	55	43,44270	17,036659	2,297223	38,83705	48,04835
2	55	35,71283	15,162917	2,044567	31,61371	39,81194
3	55	21,21693	9,077202	1,223970	18,76302	23,67084
Total	165	33,45748	16,846207	1,311475	30,86793	36,04704

Descriptives

	Minimum	Maximum
1	10,750	82,692
2	13,176	84,941
3	4,235	42,059
Total	4,235	84,941

Figura A.13: Estatísticas descritivas da freqüência extraída do corpus Florianópolis.

Test of Homogeneity of Variances

Levene Statistic	df1	df2	Sig.
8,562	2	162	,000

Figura A.14: Teste de Levene da freqüência extraída do corpus Florianópolis.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	14004,230	2	7002,115	34,862	,000
Within Groups	32538,100	162	200,852		
Total	46542,330	164			

Figura A.15: ANOVA da freqüência extraída do corpus Florianópolis.

Multiple Comparisons

	(I) Grupo	(J) Grupo	Mean Difference (I-J)	Std. Error	Sig.
Tukey HSD	1	2	7,729875*	2,702541	,013
		3	22,225774*	2,702541	,000
	2	1	-7,729875*	2,702541	,013
		3	14,495899*	2,702541	,000
	3	1	-22,225774*	2,702541	,000
		2	-14,495899*	2,702541	,000

*. The mean difference is significant at the 0.05 level.

Figura A.16: Teste de Comparações Múltiplas de Tukey da freqüência extraída do corpus Florianópolis.

Descriptives

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
1	55	1,07882E7	5,563702E6	7,502094E5	9,28412E6	1,22923E7
2	55	9,27705E6	6,984360E6	9,417709E5	7,38891E6	1,11652E7
3	55	8,92787E6	5,816294E6	7,842689E5	7,35550E6	1,05002E7
Total	165	9,66437E6	6,168407E6	4,802099E5	8,71618E6	1,06126E7

Descriptives

	Minimum	Maximum
1	3082368,750	2,658E7
2	3691266,667	4,126E7
3	1400799,412	2,561E7
Total	1400799,412	4,126E7

Figura A.17: Estatísticas descritivas da freqüência extraída do buscador 'Yahoo!'.

Test of Homogeneity of Variances

Levene Statistic	df1	df2	Sig.
,166	2	162	,848

Figura A.18: Teste de Levene da freqüência extraída do buscador 'Yahoo!'.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1,075E14	2	5,377E13	1,421	,245
Within Groups	6,133E15	162	3,786E13		
Total	6,240E15	164			

Figura A.19: ANOVA da freqüência extraída do buscador 'Yahoo!'.

Multiple Comparisons

	(I) Grupo	(J) Grupo	Mean Difference (I-J)	Std. Error	Sig.
Tukey HSD	1	2	1,511147E6	1,173264E6	,404
		3	1,860328E6	1,173264E6	,255
	2	1	-1,511147E6	1,173264E6	,404
		3	3,491809E5	1,173264E6	,952
	3	1	-1,860328E6	1,173264E6	,255
		2	-3,491809E5	1,173264E6	,952

Figura A.20: Teste de Comparações Múltiplas de Tukey da freqüência extraída do buscador 'Yahoo!'.

ANEXO B – DEFINIÇÃO DOS GRAFOS DOS GRUPOS

As figuras B.1, B.2 e B.3, apresentam as definições dos grafos G1, G2 e G3 respectivamente, no formato aceito pelo *software* de plotagem Pajek (BATAGELJ & MRVAR, 2009). Cada figura inicia apresentando a quantidade de vértices, seguida de uma listagem associando um nome ao ordinal de cada vértice. A seguir, são apresentadas as arestas, no formato de uma tripla: 'ordinal do vértice de origem', 'ordinal do vértice de destino' e 'peso da aresta'. Uma vez que estamos desconsiderando o peso nesta pesquisa, todas as arestas possuem o valor '1' na terceira posição da tripla. Devido à grande quantidade de arestas, elas foram divididas em colunas quando apresentadas nas figuras. No arquivo original, as triplas estão posicionadas uma imediatamente abaixo da outra.

```
*Vertices 22
1 "tirar"
2 "rasgar"
3 "descascar"
4 "quebrar"
5 "esmagar"
6 "estragar"
7 "fazer"
8 "montar"
9 "brincar"
10 "explodir"
11 "estourar"
12 "abrir"
13 "picar"
14 "destruir"
15 "pelar"
16 "cortar"
17 "bater"
18 "apertar"
19 "serrar"
20 "desmontar"
```

```
21 "despedaçar"
22 "amassar"
*Edges
```

8 4 1	21 5 1	5 22 1	2 11 1	16 22 1	18 6 1	22 11 1	21 19 1
7 10 1	19 21 1	12 16 1	3 7 1	1 9 1	8 7 1	17 22 1	1 4 1
13 2 1	3 16 1	17 14 1	5 2 1	4 16 1	14 9 1	4 5 1	21 14 1
11 18 1	5 11 1	7 17 1	11 10 1	2 4 1	20 7 1	3 5 1	17 7 1
1 21 1	4 9 1	14 1 1	7 5 1	3 14 1	18 11 1	6 10 1	22 1 1
17 10 1	9 7 1	13 3 1	6 5 1	5 3 1	16 4 1	14 3 1	4 14 1
1 8 1	17 16 1	2 22 1	12 5 1	4 1 1	14 16 1	16 21 1	2 14 1
16 2 1	6 12 1	18 3 1	11 5 1	3 1 1	21 2 1	6 7 1	3 12 1
14 18 1	5 6 1	16 14 1	17 5 1	6 14 1	2 5 1	11 7 1	10 11 1
2 7 1	11 14 1	21 4 1	6 18 1	4 22 1	1 13 1	7 22 1	6 16 1
22 3 1	10 4 1	19 16 1	1 19 1	11 16 1	22 5 1	6 20 1	4 20 1
22 7 1	3 22 1	1 15 1	16 11 1	7 9 1	4 10 1	13 6 1	11 2 1
4 8 1	7 1 1	4 12 1	14 7 1	7 18 1	9 4 1	11 22 1	3 18 1
9 6 1	5 17 1	2 16 1	13 19 1	19 2 1	10 7 1	18 4 1	18 7 1
2 12 1	12 1 1	10 5 1	20 9 1	18 16 1	7 6 1	16 3 1	5 21 1
3 6 1	10 17 1	6 2 1	21 13 1	17 18 1	6 4 1	15 1 1	18 5 1
6 3 1	17 1 1	4 18 1	9 8 1	13 21 1	11 6 1	22 17 1	16 18 1
5 1 1	6 22 1	11 4 1	3 4 1	9 20 1	8 1 1	21 3 1	7 20 1
11 3 1	20 14 1	8 9 1	12 4 1	19 13 1	7 3 1	20 1 1	14 2 1
7 8 1	18 14 1	1 22 1	6 11 1	1 16 1	6 17 1	19 3 1	13 16 1
12 6 1	14 11 1	17 11 1	5 7 1	16 12 1	13 1 1	18 17 1	19 4 1
17 4 1	1 2 1	18 22 1	16 17 1	22 2 1	12 3 1	16 6 1	19 1 1
7 11 1	22 16 1	20 8 1	7 2 1	4 11 1	11 17 1	14 22 1	14 4 1
6 19 1	4 13 1	19 6 1	6 8 1	3 11 1	1 20 1	5 10 1	1 14 1
8 20 1	2 13 1	16 1 1	5 18 1	16 13 1	17 3 1	3 2 1	22 4 1
2 21 1	3 13 1	14 17 1	8 14 1	9 14 1	14 6 1	6 13 1	4 3 1
14 8 1	5 16 1	21 1 1	6 21 1	6 1 1	21 12 1	4 21 1	9 1 1
20 4 1	4 2 1	2 6 1	12 21 1	4 17 1	20 6 1	3 21 1	2 3 1
1 5 1	4 19 1	1 12 1	22 14 1	17 6 1	2 19 1	7 4 1	8 6 1
16 5 1	3 19 1	22 6 1	22 18 1	3 17 1	1 7 1	12 2 1	5 4 1
14 21 1	7 14 1	4 7 1	21 16 1	13 4 1	1 17 1	16 19 1	10 6 1
							6 9 1

Figura B.1: Definição do grafo G1.

*Vertices 25
1 "tirar"
2 "arrancar"
3 "desperdicar"
4 "puxar"
5 "rasgar"
6 "desmanchar"
7 "descascar"
8 "martelar"
9 "quebrar"
10 "esmagar"
11 "fazer"
12 "montar"
13 "recortar"
14 "raspar"
15 "espedacar"
16 "estourar"
17 "abrir"
18 "cortar"
19 "bater"
20 "apertar"
21 "partir"
22 "serrar"
23 "desmontar"
24 "despedacar"
25 "amassar"
*Edges

15 18 1	11 23 1	9 13 1	15 10 1	23 2 1	20 25 1
7 2 1	4 7 1	23 11 1	6 24 1	4 5 1	21 24 1
11 18 1	2 7 1	5 2 1	21 18 1	3 5 1	20 10 1
1 21 1	9 7 1	13 5 1	2 21 1	9 5 1	2 23 1
1 19 1	23 7 1	6 5 1	25 10 1	24 5 1	18 10 1
14 5 1	13 19 1	11 5 1	18 3 1	4 2 1	19 15 1
2 18 1	7 1 1	17 5 1	22 21 1	9 24 1	1 14 1
18 21 1	6 15 1	6 18 1	7 21 1	1 23 1	23 6 1
14 18 1	11 15 1	18 2 1	24 18 1	7 6 1	4 9 1
19 20 1	9 19 1	25 16 1	22 2 1	12 24 1	3 1 1
1 11 1	10 20 1	15 13 1	9 11 1	1 17 1	2 9 1
5 3 1	8 9 1	19 13 1	24 3 1	16 9 1	24 1 1

9 6 1	15 9 1	1 6 1	23 15 1	15 1 1	23 9 1
6 9 1	14 1 1	8 19 1	5 14 1	21 9 1	5 4 1
5 1 1	7 9 1	19 1 1	1 24 1	19 9 1	14 7 1
11 9 1	20 9 1	2 1 1	6 1 1	1 4 1	11 12 1
10 1 1	2 22 1	1 9 1	11 1 1	2 3 1	11 3 1
9 25 1	18 9 1	23 1 1	10 25 1	9 15 1	
12 6 1	25 9 1	9 4 1	15 24 1	10 16 1	
15 19 1	23 24 1	24 6 1	7 4 1	9 2 1	
12 23 1	21 4 1	6 11 1	18 6 1	24 12 1	
1 18 1	19 16 1	5 13 1	25 20 1	6 7 1	
17 23 1	1 15 1	18 19 1	15 11 1	4 21 1	
7 18 1	24 23 1	5 18 1	18 11 1	11 7 1	
19 10 1	7 14 1	17 2 1	18 24 1	9 21 1	
1 5 1	3 2 1	13 21 1	21 2 1	5 24 1	
14 21 1	9 10 1	6 21 1	2 5 1	10 18 1	
21 5 1	5 7 1	7 5 1	1 7 1	16 19 1	
9 16 1	3 7 1	18 7 1	24 21 1	6 23 1	
5 11 1	24 7 1	16 10 1	23 5 1	7 11 1	
10 15 1	4 18 1	22 18 1	9 8 1	13 15 1	
9 3 1	13 18 1	7 24 1	23 12 1	18 15 1	
24 11 1	15 23 1	1 3 1	21 7 1	25 11 1	
6 12 1	7 3 1	2 4 1	10 13 1	21 14 1	
5 6 1	20 19 1	5 9 1	11 6 1	4 14 1	
9 22 1	18 13 1	4 1 1	16 25 1	19 25 1	
13 1 1	17 11 1	3 9 1	13 9 1	9 12 1	
5 17 1	15 6 1	10 9 1	20 16 1	12 11 1	
12 1 1	1 13 1	9 1 1	12 9 1	3 11 1	
11 25 1	21 22 1	24 9 1	11 17 1	24 15 1	
17 1 1	2 17 1	14 4 1	18 1 1	19 18 1	
11 24 1	23 17 1	21 6 1	19 8 1	11 2 1	
18 14 1	22 9 1	9 20 1	18 22 1	9 18 1	
7 23 1	21 1 1	5 23 1	18 4 1	5 21 1	
3 18 1	3 24 1	12 15 1	21 13 1	18 5 1	
1 2 1	1 12 1	10 19 1	9 23 1	25 19 1	
13 10 1	2 11 1	16 20 1	1 10 1	15 12 1	

Figura B.2: Definição do grafo G2.

*Vertices 31
1 "repartir"
2 "tirar"
3 "arrancar"
4 "separar"
5 "esmigalhar"
6 "fatiar"
7 "dividir"
8 "rasgar"
9 "desmanchar"
10 "martelar"
11 "despir"
12 "quebrar"
13 "esmagar"
14 "retirar"
15 "desabotoar"
16 "descosturar"
17 "descascar"
18 "furar"
19 "raspar"
20 "dar"
21 "estourar"
22 "ralar"
23 "picar"
24 "cortar"
25 "bater"
26 "partir"
27 "esfarelar"
28 "serrar"
29 "desmontar"
30 "despedacar"
31 "amassar"
*Edges

25 31 1	14 11 1	24 22 1	24 19 1	24 2 1	7 4 1
13 2 1	18 20 1	22 6 1	4 16 1	30 26 1	6 26 1
13 12 1	19 26 1	4 7 1	2 4 1	9 27 1	30 13 1
8 4 1	27 2 1	2 11 1	5 9 1	5 30 1	12 2 1
7 29 1	9 13 1	9 7 1	4 1 1	12 8 1	31 20 1
20 5 1	30 27 1	24 7 1	29 9 1	19 2 1	26 8 1

26 13 1	4 2 1	5 2 1	7 3 1	15 2 1	17 19 1
25 5 1	3 12 1	29 7 1	8 12 1	17 6 1	14 12 1
27 12 1	8 24 1	9 26 1	6 24 1	7 9 1	28 24 1
26 2 1	5 27 1	12 10 1	21 18 1	6 17 1	26 12 1
4 8 1	1 12 1	12 5 1	12 20 1	21 25 1	24 17 1
2 12 1	16 2 1	8 2 1	27 26 1	12 9 1	4 14 1
24 8 1	10 20 1	7 12 1	26 24 1	15 8 1	2 14 1
14 2 1	17 14 1	31 21 1	21 13 1	13 20 1	9 12 1
7 8 1	31 13 1	20 18 1	2 26 1	12 30 1	24 30 1
5 20 1	13 9 1	22 2 1	26 5 1	2 19 1	29 4 1
19 17 1	12 17 1	17 24 1	12 13 1	1 7 1	15 4 1
22 19 1	2 22 1	12 27 1	28 26 1	30 24 1	5 26 1
8 7 1	26 17 1	13 21 1	2 15 1	13 31 1	11 2 1
6 19 1	23 24 1	16 8 1	2 13 1	3 26 1	27 9 1
21 31 1	3 17 1	23 30 1	24 3 1	2 8 1	5 21 1
26 27 1	2 27 1	13 30 1	30 23 1	26 3 1	20 21 1
7 26 1	26 6 1	19 22 1	3 14 1	24 28 1	1 24 1
17 26 1	24 23 1	26 4 1	10 12 1	25 20 1	25 21 1
21 5 1	4 12 1	5 12 1	9 30 1	9 5 1	12 1 1
27 13 1	28 7 1	3 8 1	5 25 1	5 31 1	8 1 1
24 26 1	2 16 1	9 4 1	4 17 1	4 26 1	20 10 1
4 3 1	24 4 1	24 6 1	31 25 1	12 3 1	20 12 1
2 3 1	5 13 1	1 3 1	12 14 1	17 12 1	13 27 1
26 19 1	4 29 1	30 9 1	26 30 1	12 24 1	14 4 1
4 24 1	9 29 1	27 30 1	8 15 1	26 28 1	26 7 1
11 14 1	31 5 1	12 4 1	14 3 1	17 22 1	23 6 1
7 1 1	30 5 1	3 7 1	22 26 1	22 17 1	30 6 1
30 12 1	8 3 1	17 2 1	1 26 1	6 2 1	4 9 1
20 25 1	8 16 1	13 5 1	16 4 1	2 30 1	3 1 1
6 23 1	30 2 1	20 13 1	22 24 1	1 4 1	2 9 1
15 16 1	2 17 1	6 30 1	13 26 1	26 9 1	24 1 1
18 21 1	14 17 1	25 13 1	2 5 1	4 15 1	12 7 1
21 20 1	13 25 1	7 28 1	17 3 1	19 24 1	7 24 1
26 22 1	3 24 1	1 8 1	3 4 1	9 2 1	19 6 1
16 15 1	2 6 1	12 26 1	3 2 1	24 12 1	20 31 1
8 26 1	26 1 1	17 4 1	27 5 1	2 24 1	6 22 1

Figura B.3: Definição do grafo G3.

APÊNDICE C – ESCORES LINGÜÍSTICOS DOS VERBOS

Tabela C.1: Polissemias associadas aos verbos.

Verbo	PolWord	PolHou	Verbo	PolWord	PolHou
abrir	18	51	estourar	4	23
amassar	4	9	estragar	9	13
apertar	13	30	explodir	2	4
arrancar	11	15	fatiar	1	2
bater	19	53	fazer	23	65
brincar	3	16	furar	6	16
cortar	19	27	martelar	5	9
dar	50	68	montar	6	25
desabotoar	2	6	partir	16	24
descascar	4	7	pelar	5	2
descosturar	1	5	picar	13	33
desmanchar	9	20	puxar	18	34
desmontar	5	8	quebrar	22	22
despedaçar	3	2	ralar	3	6
desperdiçar	2	2	rasgar	9	14
despir	5	8	raspar	9	8
destruir	13	7	recortar	7	7
dividir	13	13	repartir	6	9
esfarelar	5	3	retirar	16	15
esmagar	6	7	separar	24	12
esmigalhar	3	3	serrar	0	4
espedaçar	2	2	tirar	32	53

Tabela C.2: Freqüências associadas aos verbos.

Verbo	FreqFlorian	FreqYahoo	Verbo	FreqFlorian.	FreqYahoo
abrir	51	32100000	estourar	0	818000
amassar	18	309000	estragar	40	1510000
apertar	27	2200000	explodir	0	1610000
arrancar	0	1800000	fatiar	0	114000
bater	93	10600000	fazer	491	219000000
brincar	36	6940000	furar	7	986000
cortar	25	8170000	martelar	0	89100
dar	229	89200000	montar	2	12200000
desabotoar	0	13900	partir	1	132000000
descascar	0	370000	pelar	9	35900
descosturar	0	1090	picar	0	1900000
desmanchar	0	254000	puxar	13	2990000
desmontar	0	859000	quebrar	59	6650000
despedaçar	0	45200	ralar	0	262000
desperdiçar	0	645000	rasgar	11	586000
despir	0	267000	raspar	0	410000
destruir	0	5260000	recortar	0	425000
dividir	0	7480000	repartir	0	429000
esfarelar	0	17600	retirar	1	13400000
esmagar	0	361000	separar	2	17400000
esmigalhar	0	7490	serrar	0	97800
espedaçar	0	731	tirar	107	28100000

Tabela C.3: Complexidade sintática associada aos verbos. Apenas a complexidade sintática extraída das frases dos adultos foi utilizada nos experimentos.

Verbo	Adultos	Crianças	Verbo	Adultos	Criança
abrir	0	0	estourar	20	20
amassar	0	0	estragar	0	0
apertar	0	0	explodir	20	20
arrancar	20	20	fatiar	20	20

bater	0	0	fazer	0	1
brincar	0	0	furar	0	0
cortar	1	0	martelar	20	20
dar	1	0	montar	0	20
desabotoar	20	20	partir	0	20
descascar	20	20	pelar	0	0
descosturar	20	20	picar	20	20
desmanchar	20	20	puxar	0	0
desmontar	20	20	quebrar	0	0
despedaçar	20	20	ralar	20	20
desperdiçar	20	20	rasgar	1	0
despir	20	20	raspar	20	20
destruir	20	20	recortar	20	20
dividir	20	20	repartir	20	20
esfarelar	20	20	retirar	1	20
esmagar	20	20	separar	3	0
esmigalhar	20	20	serrar	20	20
espedaçar	20	20	tirar	1	0

Tabela C.4: Posições dos verbos em cada fator lingüístico e resultado nos fatores combinados. Valores de G2.

Verbo	Posição FreqYahoo	Posição PolWord	Posição PolHou	Valor Final Combinação FreqYahoo com PolWord	Valor Final Combinação FreqYahoo com PolHou
abrir	23	12	16	17,5	19,5
amassar	6	4	5	5	5,5
apertar	16	10	14	13	15
arrancar	15	9	7	12	11
bater	20	13	17	16,5	18,5
cortar	19	13	13	16	16
descascar	8	4	3	6	5,5
desmanchar	5	8	8	6,5	6,5
desmontar	14	5	4	9,5	9

despedaçar	2	3	1	2,5	1,5
desperdiçar	12	2	1	7	6,5
esmagar	7	6	3	6,5	5
espedaçar	1	2	1	1,5	1
estourar	13	4	10	8,5	11,5
fazer	25	15	18	20	21,5
martelar	3	5	5	4	4
montar	21	6	12	13,5	16,5
partir	24	11	11	17,5	17,5
puxar	17	12	15	14,5	16
quebrar	18	14	9	16	13,5
rasgar	11	8	6	9,5	8,5
raspar	9	8	4	8,5	6,5
recortar	10	7	3	8,5	6,5
serrar	4	1	2	2,5	3
tirar	22	16	17	19	19,5

Tabela C.5: Posições dos verbos em cada fator lingüístico e resultado nos fatores combinados. Valores de G3.

Verbo	Posição FreqYahoo	Posição PolWord	Posição PolHou	Valor Final Combinação FreqYahoo com PolWord	Valor Final Combinação FreqYahoo com PolHou
amassar	12	5	8	8,5	10
arrancar	21	9	12	15	16,5
bater	26	12	20	19	23
cortar	25	12	18	18,5	21,5
dar	30	16	21	23	25,5
desabotoar	3	3	5	3	4
descascar	14	5	6	9,5	10
descosturar	1	2	4	1,5	2,5
desmanchar	9	8	14	8,5	11,5
desmontar	19	6	7	12,5	13
despedaçar	5	4	1	4,5	3

despir	11	6	7	8,5	9
dividir	24	10	10	17	17
esfarelar	4	6	2	5	3
esmagar	13	7	6	10	9,5
esmigalhar	2	4	2	3	2
estourar	18	5	16	11,5	17
fatiar	8	2	1	5	4,5
furar	20	7	13	13,5	16,5
martelar	6	6	8	6	7
partir	31	11	17	21	24
picar	22	10	19	16	20,5
quebrar	23	13	15	18	19
ralar	10	4	5	7	7,5
rasgar	17	8	11	12,5	14
raspar	15	8	7	11,5	11
repartir	16	7	8	11,5	12
retirar	27	11	12	19	19,5
separar	28	14	9	21	18,5
serrar	7	1	3	4	5
tirar	29	15	20	22	24,5

APÊNDICE D – DISTRIBUIÇÃO DOS ESCORES

Gráficos das distribuições de valores dos escores dentro de cada grupo.

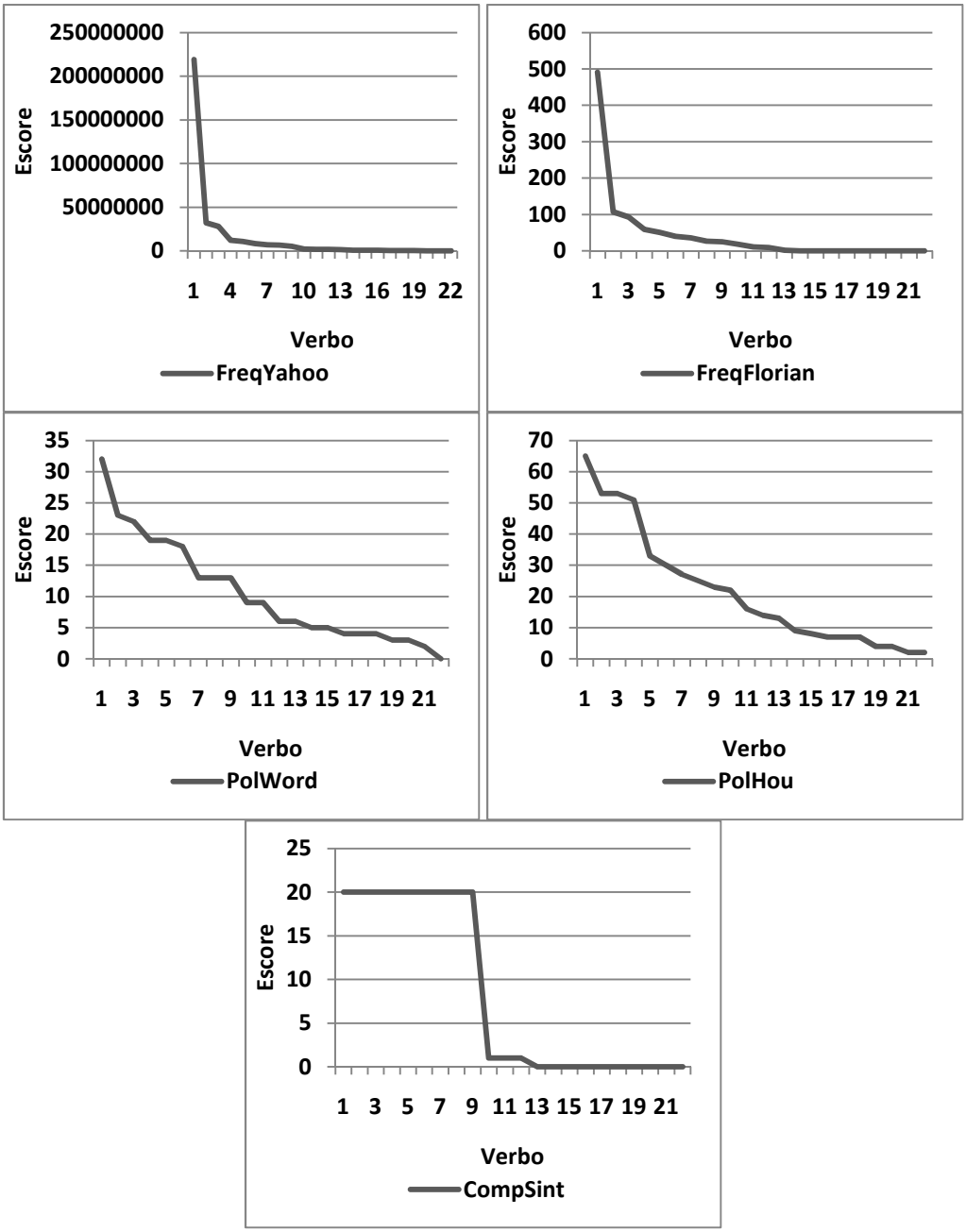


Figura D.1: Distribuição dos escores nos verbos do grafo G1 (ordem decrescente).

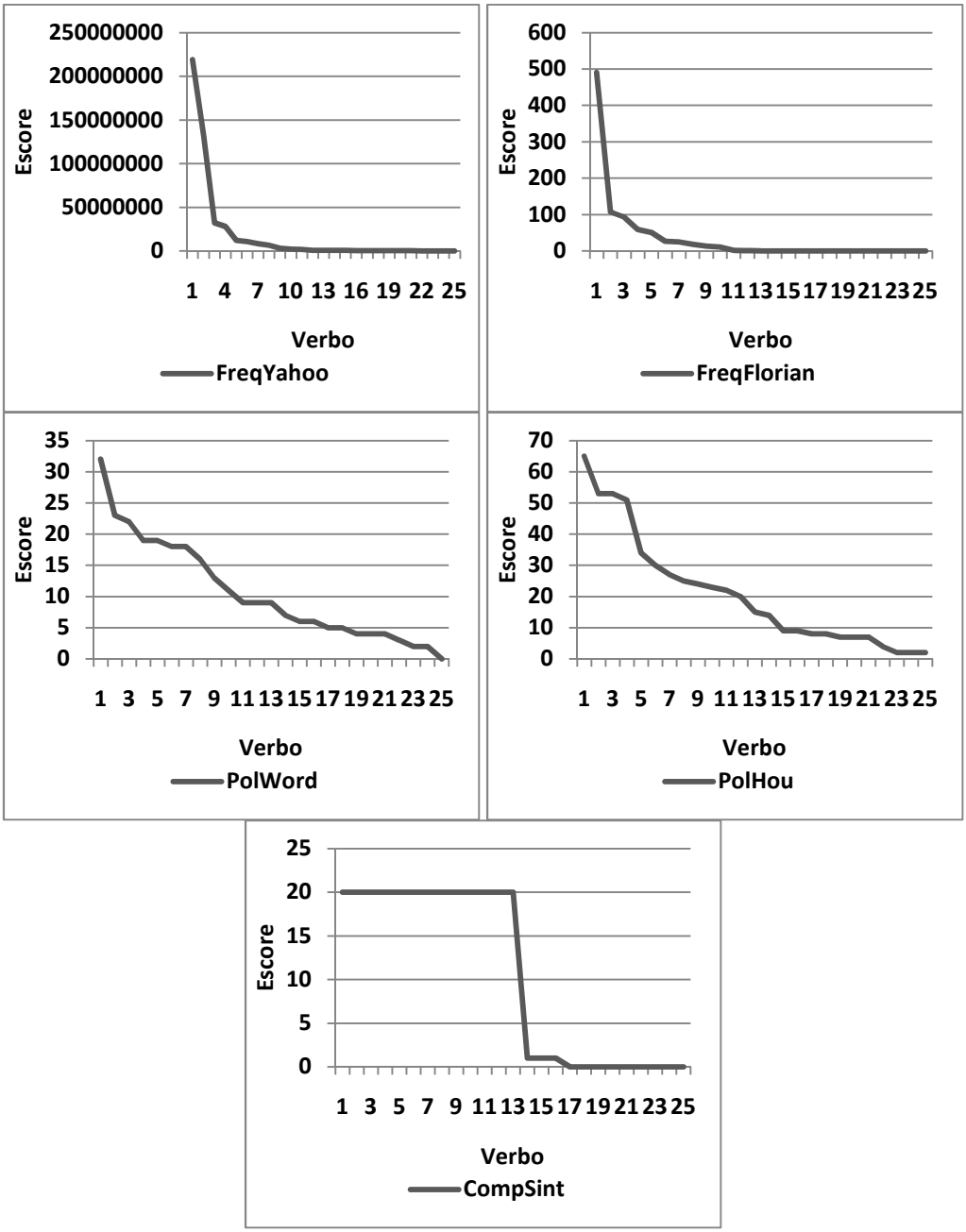


Figura D.2: Distribuição dos escores nos verbos do grafo G2 (ordem decrescente).

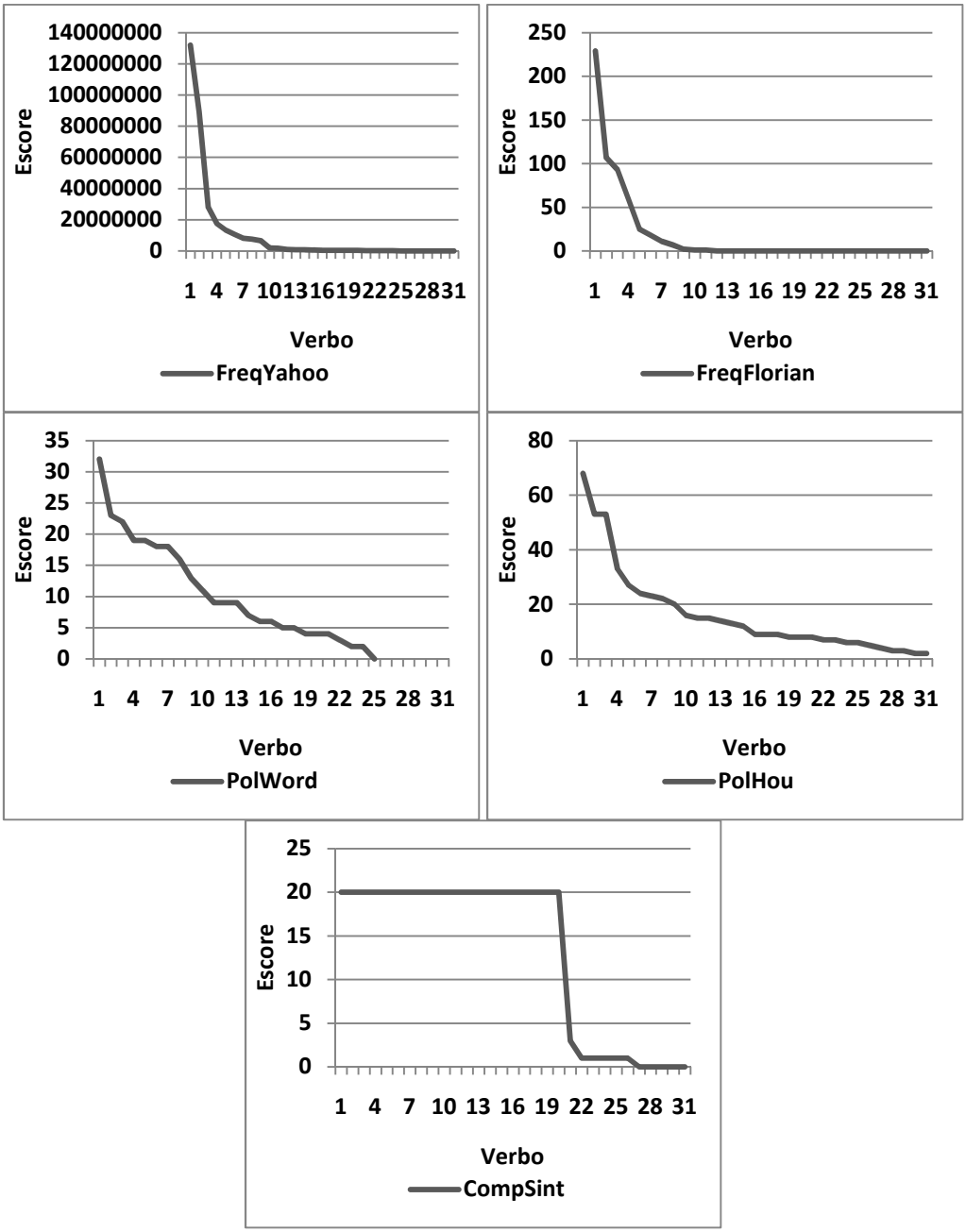


Figura D.3: Distribuição dos escores nos verbos do grafo G3 (ordem decrescente).