Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Avaliando Tradução Automática e Simplificação Textual com Redes Complexas

Diego R. Amancio
Maria das Graças Volpe Nunes

NILC-TR-09-09

Dezembro, 2009

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Contextualização e Motivação

Os conceitos e metodologias de redes complexas vêm sendo usados numa enorme variedade de áreas (Costa et al., 2008), incluindo a análise de textos escritos no que se denomina processamento de línguas naturais (PLN) (Bates, 1995). Exemplos de ferramentas de PLN mais comuns são os tradutores automáticos (Hutchins e Somers, 1992), revisores gramaticais (Martins et al., 1998) e sumarizadores automáticos (Marcu, 2000). A adequação de redes complexas nesse tipo de análise foi demonstrada em várias instâncias, a partir da comprovação de que um texto pode ser representado por uma rede livre de escala (Cancho e Solé, 2001), (Sigman e Cecchi, 2002), (Miller, 1985), (Motter et al., 2002). Resultados consolidados incluem a análise automática de qualidade de textos (Antiqueira et al., 2007), a avaliação da qualidade de sumarizadores automáticos e de estratégias de sumarização (Antiqueira et al., 2009) e a avaliação da qualidade de tradução automática (Amancio et al., 2008).

Tendo em vista o sucesso da modelagem de textos como redes complexas, este projeto apóia-se em tais resultados para desenvolver a avaliação de qualidade e complexidade de duas tarefas importantes que envolvem geração de texto em língua natural: a tradução automática, tão antiga[1] quanto a própria área de PLN, e a simplificação textual, muito recente[2] e que faz uso massivo de diversas ferramentas e de recursos de PLN sotisficados.

1.2 Objetivos

Como citado anteriormente, este trabalho divide-se em duas partes quanto à aplicação de redes complexas no processamento de línguas naturais. Na primeira parte, pretende-se mostrar que características intrínsecas aos grafos resultantes da modelagem utilizada permitem a distinção da qualidade para tradutores automáticos. Esta distinção se torna possível através da observação das dependências existentes entre a fonte da tradução e as métricas desenvolvidas na teoria dos grafos e nas recentes pesquisas com redes complexas. Na segunda parte, pretende-se alcançar resultados análogos aos das traduções, isto é, a descoberta de padrões de distinção entre textos, neste caso para diferenciar diferentes graus de simplificação em que um texto se encontra. Complementando tal estudo com simplificação textual, pretende-se também evidenciar novas estratégias de auxílio à sumarização extrativa voltada à simplificação textual.

[1] A tarefa de tradução automática iniciou-se na década de 40. Trata-se da primeira aplicação não numérica dentro da área da computação.

[2] A tarefa de simplificação textual teve início em meados da década de 90.

1.3 Organização do Trabalho

Esta monografia está organizada da seguinte maneira. Uma introdução ao tema da tradução automática e simplificação textual é apresentada nas seções 2 e 3 respectivamente. As redes complexas e suas principais métricas são introduzidas respectivamente na Seção 4 e na Seção 5, seguida de uma revisão bibliográfica envolvendo os temas de PLN e redes complexas na Seção 6. A metodologia utilizada na representação de textos como redes complexas é apresentada na Seção 7, as contribuições científicas e resultados obtidos para as duas aplicações são discutidas na Seção 8 e, por fim, a conclusão do trabalho é feita na Seção 9, com as perspectivas para as pesquisas futuras na área.

2 Tradução Automática (TA)

2.1 Visão Geral

Recentemente, com a popularização dos computadores pessoais, os avanços de *hardware*, o crescimento das redes de computadores e da *web* e o grande volume de informações disponíveis sobre uma infinidade de línguas, a tarefa de tradução automática se tornou uma das principais áreas de pesquisa em PLN. Diversas novas técnicas surgiram, principalmente estatísticas, avançando o estado da arte consideravelmente. Entretanto, os resultados ainda estão distantes do ideal, com diversos problemas a serem abordados: a gramaticalidade das sentenças (isto é, se a sentença traduzida segue as normas gramaticais da língua-alvo), a reorganização estrutural dos componentes das sentenças e do próprio texto (por exemplo, em português, o sujeito de uma sentença normalmente aparece antes do verbo, enquanto, em japonês, o verbo aparece antes), a desambiguação do sentido das palavras para a determinação da tradução correta (por exemplo, como traduzir para o português a palavra do inglês *get*, que tem vários sentidos), o tratamento de elementos co-referenciais (isto é, elementos lingüísticos que se referem a entidades introduzidas anteriormente no texto), elipses (ou seja, a omissão de termos das sentenças) e outros fenômenos lingüísticos, a tradução adequada de expressões idiomáticas e, de igual relevância, a questão da avaliação das traduções geradas pelos sistemas.

Como conseqüência, o desenvolvimento de sistemas completamente automatizados, que consideram questões lingüísticas e extralingüísticas de forma profunda, principalmente em domínios abertos ou línguas naturais irrestritas, após mais de 50 anos de pesquisa, ainda é um desafio para a área de tradução automática. De fato, ainda hoje alcançam resultados mais práticos e significativos os sistemas de TA desenvolvidos em contextos limitados, com linguagens estilizadas, regulares e específicas. Porém, sistemas baseados em sublínguas não constituem interesse para a tradução entre falantes de duas línguas naturais, por serem altamente restritos pela comunidade de uso.

Em domínios abertos, por outro lado, geralmente os textos traduzidos são compreensíveis, mas nem sempre gramaticais e raramente fluentes, implicando a necessidade de revisão humana na fase de pós-processamento.

Alguns sistemas de TA servem de auxiliares para tradutores humanos, no sentido de que realizam uma pré-tradução do texto, a ser editada ou refinada pelos tradutores humanos, a exemplo dos tradutores *Trados Workbench*[3] , *IBM Translation Manager* e *Déjavu*[4]. Outros, ainda, consideram a pré-edição do documento original, de modo a apresentá-lo em uma linguagem mais simples, como a usada pela *Xerox* no *Systran*[5], criado inicialmente para traduzir seus manuais técnicos em várias línguas.

Sistemas de TA que consideram alguma forma de edição humana, seja ela feita previamente, durante a tradução, ou posteriormente, são chamados de *Human-Aided Machine Translation*. Quando servem de auxílio à tradução humana, são chamados *Machine-Aided Human Translation*. Esses últimos incluem ferramentas de acesso a dicionários e enciclopédias, recursos de processamento de textos, verificação ortográfica e gramatical, entre outras (Boitet, 1994).

Atualmente, a *web* certamente é responsável pelo novo incentivo à TA. Com a popularização da Internet, cresceram consideravelmente a oferta e a procura de programas de TA. Diversos sistemas são capazes de traduzir páginas da Internet on-line, mensagens de correio eletrônico ou conversas via programas de *chat*.

2.2 Avaliação da Qualidade

Depois de tanto tempo, ainda não se chegou a um consenso sobre critérios e métodos de avaliação. Tradicionalmente, a avaliação tem sido feita por juízes humanos, que podem usar critérios intrínsecos para avaliar a tradução gerada (compreensibilidade, gramaticalidade, proximidade com traduções de referência feitas por humanos, etc.), ou extrínsecos, baseados na tarefa que utiliza o TA (quanto a tradução ajuda ou atrapalha a tarefa subjacente - p.e. na recuperação de informação multilíngüe). A concordância entre os juízes é medida pelo índice Kappa (Carletta, 1996), que é obtido a partir das porcentagens de concordância e discordância entre os juízes.

Tal procedimento não é adequado, no entanto, quando se trata de uma competição, como a organizada pelo NIST[6] e realizada anualmente, entre sistemas de TA. Atualmente, a avaliação desses sistemas se dá por meio de sua comparação (ocorrência de n-gramas iguais) com traduções de referência feitas por humanos. Ainda que haja um viés não automático e passível de subjetividade (será a tradução humana uma referência de fato?), a comparação é feita auto-

[3] $http://www.trados.com/$
[4] $http://www.atril.com$
[5] $http://www.systransoft.com$
[6] National Institute of Standards and Technology.

maticamente por um sistema, que fornece um índice estatístico de similaridade - índice BLEU[7] - que classifica os sistemas (Papinei et al., 2002). É preciso dizer que esta competição do NIST popularizou o BLEU e tem avançado consideravelmente o estado da arte na questão da avaliação. Neste contexto, insere-se este projeto, relacionando características de redes com qualidade, através do emprego de recursos linguísticos básicos.

3 Simplificação Textual

3.1 Visão Geral e Avaliação de Complexidade

Define-se simplificação textual como o processo de desacentuar a complexidade linguística de um texto através da redução do léxico e atenuação de construções sintáticas complexas, com nenhuma ou pouca perda da informação principal (Max, 2006). Trata-se de uma tarefa importante tanto para pessoas quanto para máquinas. No primeiro caso, pessoas com baixo nível de letramento, com deficiências cognitivas (afasia, dislexia, entre outros), crianças e aprendizes de línguas estrangeiras são beneficiadas pela apresentação simples de um texto. Já no segundo caso, a simplificação textual pode ser utilizado como um pré-processamento para outras tarefas a fim de simplificar uma etapa posterior cuja eficiência muitas vezes é dependente do grau de complexidade da entrada[8].

Existem vários tipos de simplificação, desde a sumarização (apenas para diminuir o tamanho do texto) até o aumento da inteligibilidade (eliminação de palavras pouco frequentes e estruturação sintática pouco complexa) e da legibilidade (forma como o texto se apresenta). Neste ponto, vale a pena ressaltar o tipo de simplificação utilizada no projeto. Enquanto a Seção 8.3.1 e a Seção 8.3.2 buscam encontrar padrões em simplificações com enfâse na legibilidade e inteligibilidade, a Seção 8.3.3 enfatiza um algoritmo para identificação de palavras chaves com fins de sumarização aplicada à simplificação.

Quanto à tarefa de distinção de níveis de simplificação de um texto é interessante citar trabalhos como em Maziero et al. (2002), o qual analisa a inteligibilidade de um corpus contando por exemplo o número de construções que utilizam voz passiva, número de marcadores discursivos ambíguos e outras características que são indiretamente úteis a tarefa de quantificação da complexidade de um texto. Para isto, utiliza-se o *parser* PALAVRAS (Bick, 2000), um dos principais analisadores sintáticos automáticos para o português do Brasil.

Outro índice amplamente conhecido, desta vez relacionado à legibilidade é o índice *Flesch* (φ) (Flesch, 1948), calculado segundo a equação 1.

$$\varphi = 206,835 - 1,015 \cdot \bar{\rho} - 84,6 \cdot \bar{\varsigma} \tag{1}$$

[7]BiLingual Evaluation Understudy

[8]Tradução automática, sumarização de textos, extração de informação e *parsing* são exemplos de tais tarefas.

Valor	Facilidade de Leitura
$0 \preceq \varphi \prec 40$	Muito difícil
$40 \preceq \varphi \prec 50$	Difícil
$50 \preceq \varphi \prec 60$	Razoavelmente difícil
$60 \preceq \varphi \prec 70$	Razoável
$70 \preceq \varphi \prec 80$	Razoavelmente fácil
$80 \preceq \varphi \prec 90$	Fácil
$90 \preceq \varphi \preceq 100$	Muito fácil

Tabela 1: *Interpretação do Índice de Facilidade de Leitura de Flesh*

sendo que $\bar{\rho}$ representa o número médio de palavras por sentença e $\bar{\varsigma}$ representa o número médio de sílabas por palavra do texto. Sua interpretação é dada pela Tabela 1.

Um outro valor relacionado é o índice *Flesh-Kincaid* (κ) (Kincaid et al.,1975), calculado segundo a equação 2. Seu valor pode ser interpretado como o nível de estudo americano mínimo necessário para compreensão do texto ou o número mínimo de anos de estudo necessário para o entendimento.

$$\kappa = 0,39 \cdot \bar{\rho} + 11,8 \cdot \bar{\varsigma} - 15,59 \tag{2}$$

4 Redes Complexas

Pode-se definir formalmente uma rede ou grafo como uma estrutura $G = \{V, E\}$, sendo V um conjunto de vértices ou nós e E uma relação entre tais vértices representada pelo conjunto de arestas. Apesar de representar um conceito bastante difundido na matemática discreta e na ciência da computação[9] grande parte do interesse em redes têm se dado devido ao fato de que muitos sistemas e fenômenos complexos que existem são melhor entendidos quando modelados como redes. Exemplos desses sistemas são as cadeias alimentares, a rede rodoviária de uma cidade, a *World Wide Web*, as redes neurais e as redes de relações sociais entre indivíduos[10].

Tradicionalmente, os principais métodos para compreensão das características essenciais de uma rede eram restritos a uma representação gráfica em forma de linhas e nós. No entanto, esta abordagem tornou-se inviável com a dimensão crescente das redes. Assim, as características de tais redes e suas propriedades passaram a ser analisadas em larga escala, com conseqüente mudança na modelagem tradicional, evidenciando uma crescente necessidade de entender o comportamento do sistema como um todo, movendo-se para além das abordagens reducionistas.

[9]Um dos clássicos problemas envolvendo grafos, conhecido como pontes de Königsberg de Euler, data do século XVIII.

[10]Exemplos de redes sociais são os sites myspace (www.myspace.com), facebook (www.facebook.com) e orkut (www.orkut.com). Cada usuário é um vértice desse grafo social e a relação de amizade entre usuários representam as arestas.

Esta abordagem tradicional considerava as redes como sendo de caráter aleatório, isto é, com uma característica democrática - os nós apresentam aproximadamente o mesmo número de arestas - como é possível observar na Figura 1, com sua respectiva função de probabilidade para o número de ligações de um vértice descrita na Figura 3, à esquerda. No entanto, percebeu-se algo diferente em certas redes, denominadas redes complexas. Por exemplo, esperava-se que a Internet apresentasse esta topologia randômica, no entanto verificou-se uma rede com características semelhantes à descrita na Figura 2, com apenas poucos vértices apresentando alta conectividade. Uma análise detalhada mostrou que na Internet 80% dos nós possuíam menos de 4 links e menos de 0,01% mais de 1000. A distribuição encontrada neste caso é mostrada na Figura 3, à direita. Particularmente, este comportamento adiciona uma nova nomenclatura a estas redes, denominando-as como *scale free*, ou livre de escala (Cancho e Solé, 2001), (Sigman e Cecchi, 2002), (Miller, 1985), (Motter et al., 2002).

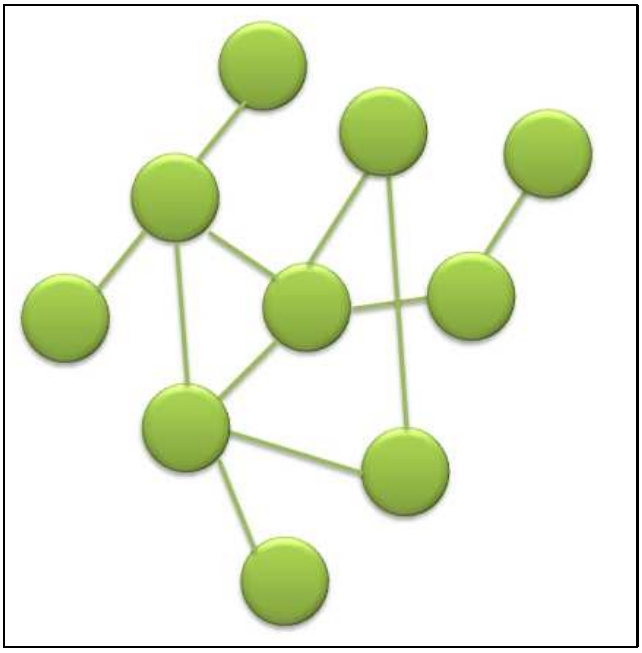

Figura 1: *Exemplo de rede aleatória.*

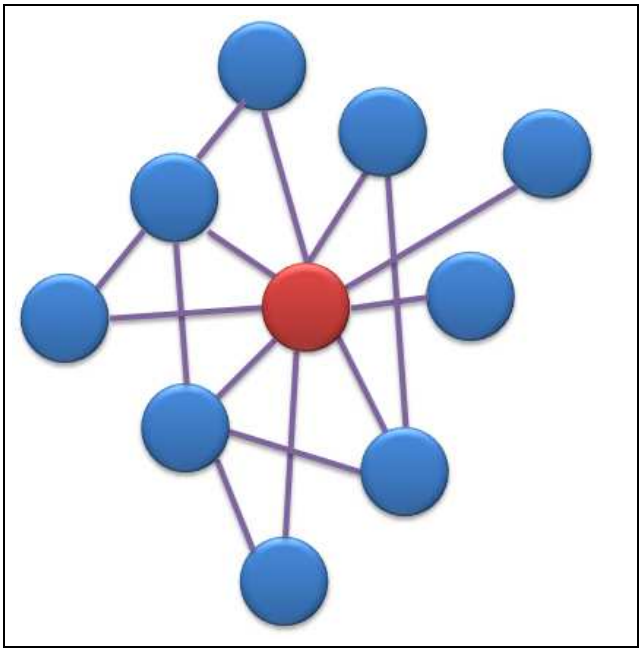

Figura 2: *Exemplo de rede complexa.*

Grande parte das recentes descobertas em novas modelagens está relacionada justamente à maneira como as redes do mundo real diferem das redes aleatórias (Newman, 2003), como exemplificado no exemplo da Internet. Uma destas diferenças é ilustrada no padrão encontrado na Figura 3 (esquerda), o qual é dito seguir uma distribuição do tipo lei de potências, mais especificamente como P(k) $\approx k^{-\gamma}$. Tal expressão significa que a probabilidade de um vértice possuir k ligações é inversamente proporcional a k elevado a certa potência. Esta distribuição leva a uma característica muito importante da rede: a presença de vértices que praticamente sustentam a rede como um todo. Tais vértices são comumente conhecidos como *hubs*. *Hubs* são importantes para a estrutura de uma rede no sentido em que proporcionam uma maior

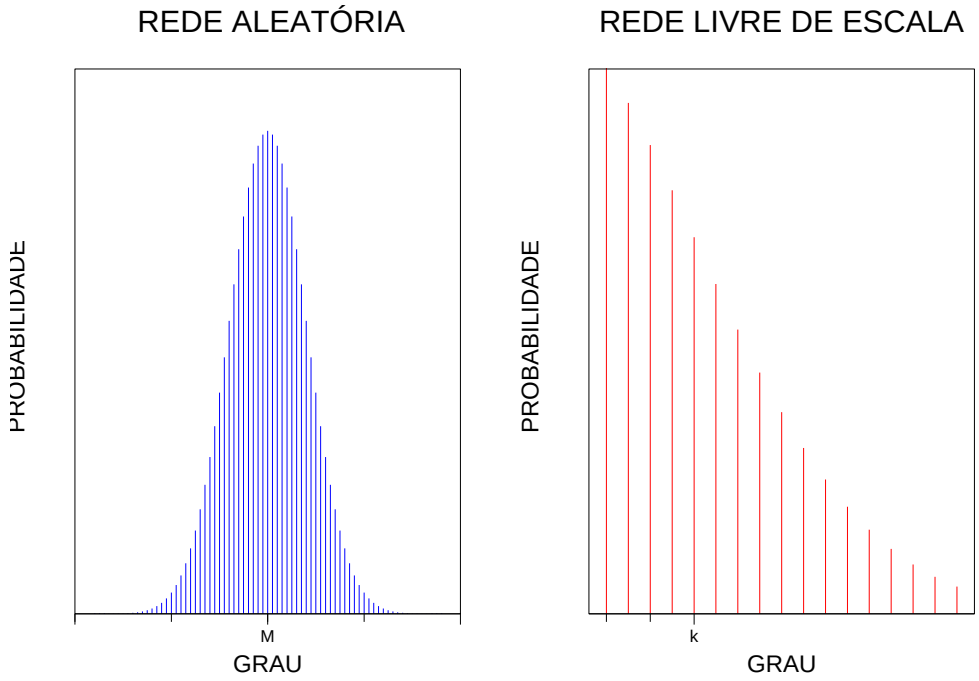

Figura 3: *Distribuição de probabilidade para rede aleatória (esquerda) e rede livre de escala (direita), com M representando o grau médio da rede aleatória e k representando um dos graus da rede livre de escala.*

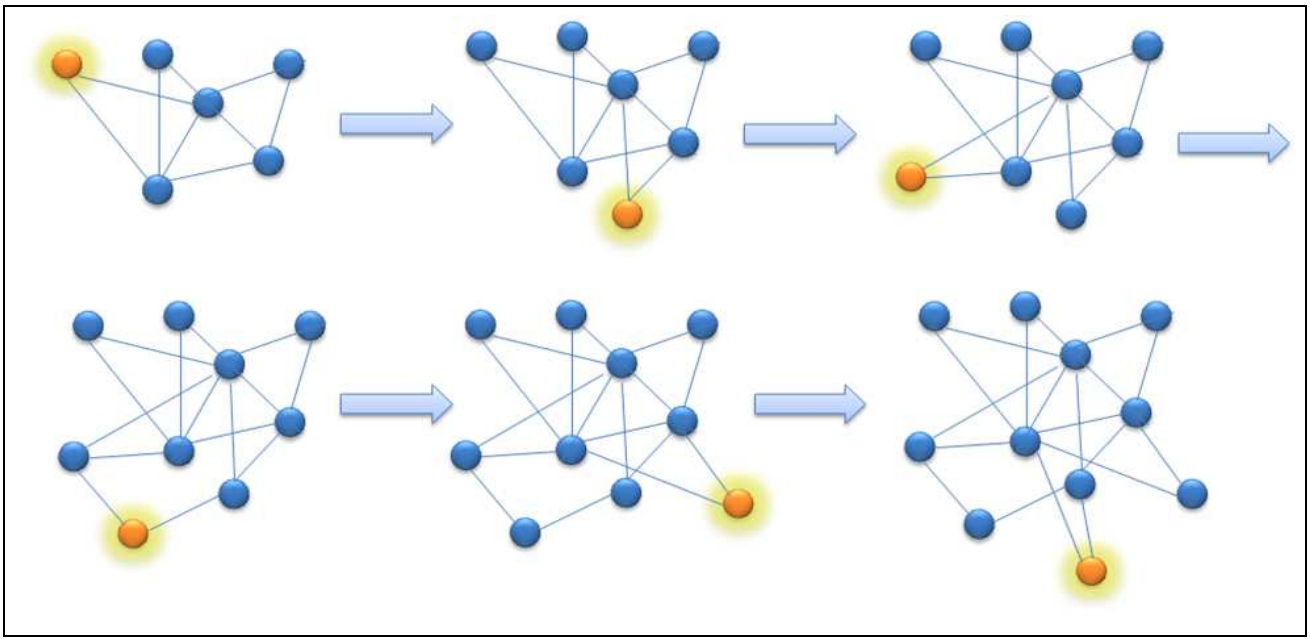

Figura 4: *Vértices com mais ligações tendem a receber novas ligações após a inserção de um novo vértice.*

confiabilidade à rede, tornando-a resistente a falhas acidentais. Isto acontece pois como ataques são investidos aleatoriamente a alguns vértices, provavelmente os *hubs* não serão atingidos - já que representam uma pequena parte da rede -, daí a estrutura da rede tende a ser não muito prejudicada, justamente devido ao fato de os *hubs* fornecerem conexões entre vários vértices da rede. Voltando ao caso da Internet, se ataques fossem feitos aleatoriamente a 80% dos vértices, esta não sofreria um colapso total devido ao suporte oferecido pelos *hubs*. Obviamente, isto não aconteceria se a rede em questão fosse uma rede aleatória. Contudo, é importante notar que um pequeno ataque - entre 5% a 15% dos vértices - centralizado nos *hubs* causaria um enorme impacto na estrutura global da rede, levando a um colapso total.

Outra característica importante em uma rede *scale free* é que vértices que se apresentam fortemente conectados têm uma tendência maior em receber novas conexões. Um exemplo da dinâmica de uma rede *scale free* é exibido na Figura 4. Isso é claramente comprovado em algumas redes que seguem a lei de potências: a Internet, atores de *Hollywood* e citações em artigos. Na Internet, sites amplamente ligados tendem a receber mais ligações (*links*) do que sites poucos ligados, geralmente desconhecidos por serem pouco acessados. Na rede de atores de *Hollywood*, atores famosos tendem a receber mais convites para atuar em novos filmes em relação a atores desconhecidos. Por fim, na rede de citações, atores bastante citados também têm maior chance de serem citados novamente em relação a pesquisadores que estão iniciando seu trabalho.

Redes *scale free* também são propensas a infecções contagiosas: o ataque a um vértice faz

com que provavelmente este contamine algum *hub*, o qual espalha para o resto da rede. No entanto, esta topologia pode gerar uma nova estratégia de vacinação, na qual bastaria vacinar os *hubs* apenas, caso fosse possível apontá-los na população.

A propriedade *scale free* e sua conseqüente estrutura de *hubs* definem praticamente a segunda grande propriedade de uma rede complexa: a característica conhecida como *small-world* (Watts, 1999), (Milgran, 1967). Uma rede é dita ser *small-world* se a distância entre quaisquer dois nós é relativamente pequena. Isto é verdade justamente devido à presença dos *hubs* que, por estarem conectados a vários vértices, tendem a diminuir a distância entre dois vértices quaisquer.

Novamente, para a propriedade *small-world*, podem-se citar vários fenômenos naturais inseridos neste tipo de comportamento: redes de proteínas, redes neurais e redes de alimentação, além de novamente a Internet (Costa et al., 2008). Especialmente, nesta última, foi analisado que embora envolva uma quantidade enorme de vértices, em média um caminho entre quaisquer dois nós deve ser percorrido por no máximo 19 nós intermediários. Por fim, é importante citar que uma rede complexa apresenta também a capacidade de *clustering* dos vértices, isto é, a tendência de aglomerados com vértices altamente conectados. Será visto que este comportamento pode ser quantificado em uma das medidas extraídas: o *cluster coefficient* ou coeficiente de aglomeração.

Enfim, pode-se concluir que o conceito de rede complexa, originalmente da física estatística (Albert e Barabási, 2002) e amplamente difundido em outras área (Costa et al., 2009) está fortemente relacionado com os fenômenos *small word*, *scale free* e *clustering*, sendo tais fenômenos responsáveis por sua organização especial cuja estrutura segue princípios complexos de organização.

4.1 Caracterização de redes complexas

Como citado anteriormente, um dos motivos do sucesso das redes é sua facilidade em modelar vários sistemas complexos de natureza distinta. No entanto, para que se possa aproveitar efetivamente as vantagens de tal modelagem é necessário caracterizar a rede quantitativamente a fim de diferenciar cada uma por suas características estruturais estáticas e dinâmicas. A idéia básica de caracterização consiste no mapeamento de uma rede complexa genérica G em um vetor descritivo, $\overrightarrow{\mu}$, sendo cada componente do vetor uma medida descritiva, conforme ilustra Figura 5.

Assim, toda rede pode ser caracterizada por um único vetor $\overrightarrow{\mu}$, que por sua vez pode ser utilizado tanto para descrições da estática como do comportamento dinâmico da rede. No primeiro caso, duas redes podem ser comparadas, por exemplo, pelo $\| \overrightarrow{\mu} \|$ ou também é possível verificar a estabilidade de uma medida a pequenas perturbações na topologia da rede. No segundo caso, pode-se comparar a evolução da rede, como por exemplo os valores possíveis do

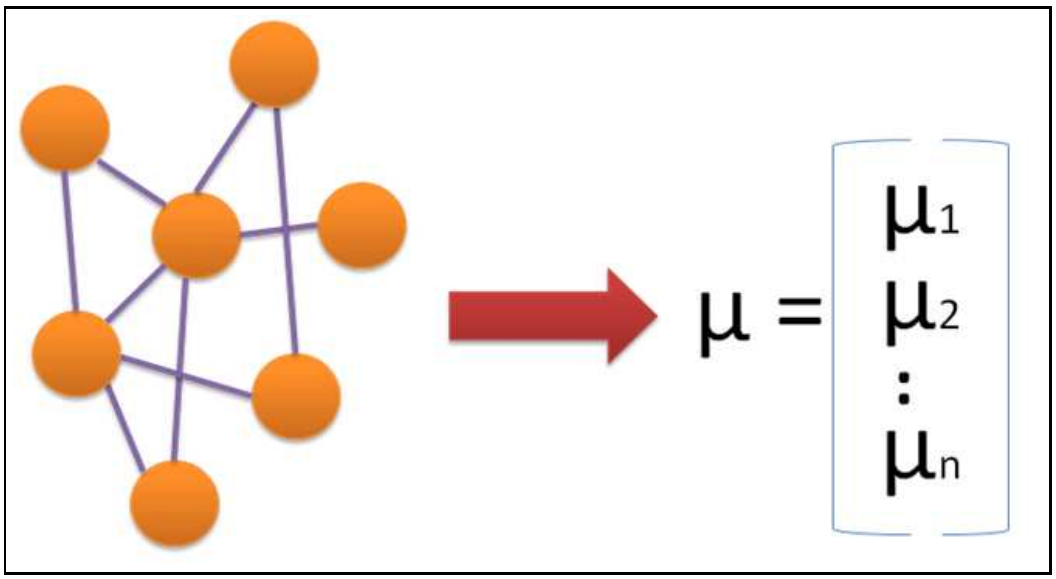

Figura 5: *Exemplo de caracterização de uma rede complexa por métricas extraídas, denominadas pelos componentes μ_i do vetor μ. O processo inverso, caracterizado pela obtenção da rede através do vetor μ é denominado representação.*

$\|\overrightarrow{\mu}\|$ a cada adição de uma aresta.

A caracterização de redes por métricas têm mostrado sua importância, no entanto, algumas questões geralmente surgem quanto ao conjunto de métricas que deve ser utilizado. Pode-se dizer que medidas simples como graus de entrada e saída, coeficiente de aglomeração e caminhos mínimos são as mais utilizadas, compondo portanto o conjunto de medidas básicas. No entanto, seriam estas suficientes para caracterizar univocamente uma rede ? Em muitos casos esta resposta é negativa, já que algumas medidas podem ser redundantes, o que faz com que redes distintas sejam mapeadas em um único conjunto de valores $\overrightarrow{\mu}$. Assim, várias outras medidas são propostas a fim de caracterizar univocamente as redes. Nesta linha de pesquisa, novas pesquisas (Costa e Andrade, 2007) tentam encontrar o melhor conjunto de medidas para caracterização e discriminação de modelos de redes complexas.

5 Métricas da rede

Esta seção apresenta o conjunto de métricas implementadas e utilizadas na caracterização de redes complexas. Destacam-se dois tipos de métricas: padrões e hierárquicas.

5.1 Medidas padrões

5.1.1 Graus de Entrada e Saída

As medidas de grau, também conhecidas como conectividade na literatura física (Dorogovtsev e Mendes, 2002) podem ser divididas em dois tipos. O grau de saída de um nó (OD) corresponde à soma das ponderações das arestas que saem de um nó. Analogamente, define-se o grau de entrada de um nó (ID), como a soma das ponderações das arestas que incidem naquele nó. Ambas as medidas globais são calculadas como a média de todos os nós da medida em questão. É importante notar que ao se trabalhar com a medida global[11] dos graus, OD e ID são sempre iguais, já que a soma de todas as arestas que saem é igual a soma de todas as arestas que incidem naquela rede. Em termos da matriz de adjacência utilizada para representar uma rede com ponderações[12] (Barthélemy et al., 2007) tem-se que para um dado nó i seu OD e seu ID são calculados pelas expressões abaixo, sendo que N representa o número total de nós[13] :

$$ID(i) = \sum_{j=1}^{N} W_{ij} \tag{3}$$

$$OD(i) = \sum_{j=1}^{N} W_{ji} \tag{4}$$

Outras medidas também usadas neste projeto são o grau máximo e mínimo da rede:

$$OD_{max} = \max_i OD_i \qquad ID_{max} = \max_i ID_i \tag{5}$$

Apesar da simplicidade da definição de tais medidas, estas têm se mostrado extremamente úteis na distinção de várias redes, como em redes lingüísticas de tradução automática (Amancio et al., 2008).

5.1.2 Coeficiente de Aglomeração

O coeficiente de aglomeração (CC) refere-se a uma das medidas de análise de estrutura cíclica da rede complexa e sua capacidade de formar conjuntos de nós conectados entre si de forma estreita, chamados de aglomeração. Existem vários tipos de coeficiente de aglomeração. Aquele utilizado neste projeto trata da relação de três vezes o número de triângulos[14] (N_t) da rede e o número de conjuntos de três nós conectados[15] (N_3) entre si, obtidos do grafo não orientado

[11] A medida global difere da medida local, pois esta última refere-se a um nó em específico, enquanto a primeira refere-se à média da medida para todos os nós da rede.

[12] A modelagem da rede como matriz de ponderações está descrita na Seção 7.

[13] Neste projeto, cada palavra distinta do texto pré-processado será modelada como um nó.

[14] Triângulos são três vértices totalmente conectados.

[15] isto é, conjunto de três vértices, cada qual podendo alcançar cada um dos outros vértices de forma imediata ou não.

correspondente. Se a_{ij} representa um elemento qualquer da matriz de adjacência do grafo, então:

$$N_t = \sum_{k>j>i} a_{ij}a_{ik}a_{jk} \tag{6}$$

$$N_3 = \sum_{k>j>i} (a_{ij}a_{ik} + a_{ji}a_{jk} + a_{ki}a_{kj}) \tag{7}$$

$$C = \frac{3N_t}{N_3} \tag{8}$$

onde a equação 8 representa a medida global de coeficiente de aglomeração. Muitas vezes também se faz necessário o uso de tal medida localmente, o qual é obtida de cada nó. Segue que, da mesma definição, obtém-se :

$$N_t(i) = \sum_{k>j} a_{ij}a_{ik}a_{jk} \tag{9}$$

$$N_3(i) = \sum_{k>j} a_{ij}a_{ik} \tag{10}$$

$$C(i) = \frac{N_t(i)}{N_3(i)} \tag{11}$$

Outra definição para a medida local, mas também envolvendo o mesmo conceito de aglomeração em uma rede pode ser obtida imaginando-se N_c como sendo a cardinalidade de um conjunto de nós que recebem arestas do vértice i. Se B é definido como o número de arestas que realmente existem neste conjunto definido[16], então o coeficiente de aglomeração também pode ser definido como:

$$C(i) = \frac{B}{N_c(N_c - 1)} \tag{12}$$

Caso Nc ≤ 1, por definição toma-se o valor de CC como zero. É possível observar que o coeficiente de aglomeração está sempre entre zero e um, diretamente desta sua definição. Uma relação geralmente feita com esta medida e redes sociais é que ao considerar vértices como pessoas e arcos como relações sociais (relações de amizade), pode-se dizer que uma pessoa tem alto coeficiente de aglomeração se vários de seus amigos são também amigos entre si.

[16]É importante observar que o termo $N_c(N_c$ -1) representa o número de arestas caso todos os vizinhos estivessem ligados entre si

5.1.3 Caminhos Mínimos

Um caminho mínimo, uma das medidas relacionadas com distâncias dentro da rede, é também conhecido como caminho geodésico entre dois nós. Trata-se de um caminho conectando dois nós com distância mínima, levando-se em consideração a soma das ponderações das arestas percorridas por este caminho. Tal caminho, que não precisa ser único, é calculado entre um dado nó fixo e todos os outros nós. Para a rede como um todo, define-se o caminho mínimo como a média dos comprimentos todos os caminhos mínimos existentes entre quaisqueir dois vértices. Localmente, define-se o caminho mínimo de um vértice como :

$$SP(i) = \frac{\sum_{j \neq i} d(i,j)}{N-1} \tag{13}$$

Três variações de caminhos mínimos são usados nos experimentos. Enquanto SP_1 trabalha com as próprias arestas, SP_2 considera o complemento das arestas (W_{max} - W(i,j) + 1) e SP_3 considera o inverso dos pesos. As duas últimas variações se diferenciam do cálculo convencional de caminho mínimo pois é dada maior importância para as arestas mais importantes do grafo, isto é, aquelas com maiores ponderações. É importante lembrar também que se não houver caminho entre dois vértices, o caminho mínimo é tomado como o número de vértices (N) da rede para SP_1 e N*W_{mean} para SP_2 e SP_3, com W_{mean} é a média dos pesos das arestas presentes na rede.

5.1.4 *Components Dynamics Deviation*

Esta medida (CDD) apresenta uma característica diferente em relação às outras medidas verificadas até aqui: tal medida analisa não a rede após estar totalmente concluída, mas sim a rede a cada associação lida, isto é, a cada adição de uma nova palavra ao texto. Mais especificamente, a cada associação lida, o número de componentes fracamente conexos[17] no grafo é anotado. Uma relação binária que associa o número de associações lidas e o número de componentes fracamente conexos é construída. Se essa relação for denotada por f_a e uma reta de referência que passa pelos pontos final e inicial do domínio de f_a for denotada por f_s, então o desvio na dinâmica de componentes é dado por :

$$CDD = \frac{\sum_{x=1}^{L} \left| f_a(x) - f_s(x) \right|}{N \cdot L} \tag{14}$$

com L representando o número total de associações de palavras e N o número total de nós na rede.

Pode-se observar que o desvio é mínimo quando a cada associação lida, o número de com-

[17]Diz-se que uma componente de um grafo é fracamente conexa se existe um caminho entre cada par de vértices do grafo não orientado resultante da componente analisada.

ponentes se reduz de uma unidade, assim a relação obtida iguala-se à reta de referência.

5.1.5 Eficiência Global

A definição padrão de distância geodésica média, apesar de bem intuitiva, apresenta a desvantagem de poder divergir caso exista na rede nós desconectados. Para superar esta situação define-se a medida relativa à eficiência global (GE) da rede na equação 15.

$$GE = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \tag{15}$$

A interpretação desta medida está relacionada com a capacidade da rede em trocar informações entre quaisquer dois nós, dado que se uma distância d_{ij} for pequena esta contribuirá de forma mais significativa em relação a uma distância d_{ij} grande. Nota-se que a fórmula acima é uma das formas que previne a divergência das medidas relacionadas com distâncias, portanto, mostra-se útil no caso do grafo apresentar mais de uma componente[18].

Pode-se ressaltar também que o inverso de GE também tem sido utilizado como medida em redes complexas, sendo conhecido também como *média harmônica das distâncias geodésicas.*

5.1.6 Vulnerabilidade da rede

Sabe-se que em uma rede complexa, nem todos os nós apresentam a mesma importância na manutenção da estrutura e funcionamento da rede como um todo. Para perceber isto, basta observar intuitivamente os *hubs*[19] e logo nota-se que estes são essenciais para a rede devido ao seu alto grau de conectividade. Mas os vértices importantes não se resumem apenas aos *hubs*: supondo um considerável subgrafo com característica muito próxima de uma árvore binária, a raiz desta árvore seria um ponto de vulnerabilidade, já que se trata de um único meio de ligação entre duas árvores e este vértice raiz não necessariamente seria um *hub*.

Para quantificar o grau de importância de um dado vértice na rede, deve-se verificar a variação da estrutura da rede quando o nó é removido da rede, isto é, todas as arestas referentes àquele vértice, além do próprio vértice são removidos. Uma maneira comum seria considerar como variação da rede uma medida de desempenho, tal como a eficiência global da rede. Assim, usualmente define-se a vulnerabilidade de um vértice i como sendo:

$$V_i = \frac{GE - GE_i}{GE} \tag{16}$$

na qual GE refere-se à eficiência global da rede com todos os nós e GE_i refere-se à eficiência global da rede após a retirada do vértice i. Nota-se que V_i representa a diferença relativa

[18] caso da rede de adjacência de palavras obtida da modelagem do texto como rede complexa.
[19] Ver seção 4.

percentual (se multiplicado por 100) refletindo a remoção daquele nó. Diferentemente das outras medidas vistas até aqui, a medida global correspondente não é a média obtida da vulnerabilidade de cada nó. Neste caso o importante é conhecer o pior da caso da rede, então segue que a vulnerabilidade da rede é:

$$V = \max_i V_i \tag{17}$$

5.1.7 *Search Information*

Muitas vezes deseja-se conhecer a dificuldade em se buscar um vértice em uma rede complexa sobre um dado caminho. Esta medida (SI) relaciona-se com a distribuição dos graus na rede e também com a entropia da rede[20]. Esta dificuldade geralmente é qualificada considerando caminhos aleatórios sobre a rede e a probabilidade de um dado caminho mínimo ser seguido aleatoriamente. Seja SP um caminho mínimo entre os vértices i e b da rede. Começando pelo nó i um caminho aleatório é tomado. Se $p(i,b)$ representa a probabilidade do caminho mínimo ser seguido, então :

$$p(i,b) = \frac{1}{k_i} \prod_{j \in SP} \frac{1}{k_j - 1} \tag{18}$$

na qual a produtória é realizada sobre todo o caminho, com exceção dos vértices final e inicial. Em uma rede não orientada, k_i é tomado como o grau do nó. Para este projeto particularmente, na qual se faz uso principalmente de grafos orientados toma-se k_i como o grau de saída do vértice i.

A medida de SI está relacionada com esta probabilidade. Mais especificamente, a medida de SI entre dois vértices i e b está relacionada com a entropia e com todos os caminhos mínimos entre estes dois vértices da seguinte forma:

$$S(i,b) = -\log_2 \sum_{SPs} p(i,b) \tag{19}$$

na qual o termo SPs representa que a soma é tomada por todos os caminhos mínimos entre os vértices i e b. Feito isto para todos os possíveis caminhos na rede, a média é utilizada como a medida final. Assim a definição final de SI é dada por :

$$SI = \frac{1}{N^2} \sum_i \sum_j S(i,j) \tag{20}$$

[20]Ver seção 5.1.11

5.1.8 Coeficiente cíclico

Para verificar se a rede de adjacências de palavras é cíclica e o quanto este característica está presente na rede, define-se o coeficiente cíclico. Trata-se de uma medida local, portanto cada vértice possui sua medida independentemente do outro. Por definição, toma-se o coeficiente cíclico como sendo:

$$Cyc_i = \frac{2}{k_i(k_i - 1)} \sum_j \sum_k \frac{a_{ij}a_{ik}}{S_{ijk}} \tag{21}$$

na qual o termo S_{ijk} representa o tamanho do ciclo mínimo que possui os nós i, j e k. O menor valor que S_{ijk} pode assumir é quando estes três vértices estão ligados em ciclo diretamente, de forma a formar um triângulo e conseqüentemente $S_{ijk} = 3$. O termo k_i representa o grau do vértice, ou seja, a soma do seu grau de entrada com seu grau de saída. Esta fórmula local representa a média dos inversos dos menores ciclos entre um nó e seu vizinho. Então esta medida será quanto maior quanto mais ciclos de menores tamanhos a rede possuir, o que está de acordo com a intuição sobre o quão cíclica é uma rede. Observe também que o termo inverso de S_{ijk} evita divergência quando não existe ciclo entre os três vértices, isto é, $S_{ijk} = \infty$. Naturalmente, para a rede como um todo, toma-se a média das medidas dos vértices.

5.1.9 *Rich Club*

O fenômeno conhecido como *Rich Club* está relacionado ao fato de que nós semelhantes apresentarem certa tendência em possuírem ligações entre si. Suponha uma rede social com classificação dos nós pela classe econômica. Sabe-se que existe uma maior tendência de haver ligação[21] de uma pessoa da classe alta com outra pessoa da classe alta, da mesma forma que pesquisadores importantes tendem a se juntar para publicar artigos de forma conjunta.

Diferentemente das outras métricas que são calculadas para cada vértice, originalmente define-se o *Rich Club* como sendo relativo a um dado grau presente em um ou mais vértices da rede. Assim, tal medida de grau k de uma rede é dada como o conjunto de nós com grau maior que k, isto é o conjunto :

$$\mathcal{R}(k) = \{v \in V(G) \mid k_v > k\} \tag{22}$$

onde V(G) representa o conjunto de vértices do grafo G. Assim, define-se a medida de grau k como:

$$\mathcal{RC}(k) = \frac{1}{|\mathcal{R}(k)|(|\mathcal{R}(k)| - 1)} \sum_{i,j \in \mathcal{R}(k)} a_{ij} \tag{23}$$

com a_{ij} representando o elemento da matriz de adjacência. A definição usada neste projeto

[21] Um exemplo de ligação poderia ser amizade/contato pessoal entre as pessoas que formam a rede social.

trata grafos ponderados orientados. Neste caso, a definição é semelhante a anterior. Tomando w como o a soma do grau de entrada e saída (soma das ponderações) e definindo-se o conjunto de grau w, $R(w)$, como sendo aquele cujos os vértices possuem w maiores que o do vértice considerado, define-se agora o análogo para grafos ponderados:

$$\mathcal{RC}(w) = \frac{\sum_{i,j \in \mathcal{R}(w)} M_{ij}}{\sum_{i \in \mathcal{R}(w)} w_i} \tag{24}$$

na qual M_{ij} representa o elemento da matriz ponderada do texto modelado como rede. Neste projeto, toma-se como medida de *Rich Club* referente ao grau médio da rede:

$$RC \equiv \mathcal{RC}\left(\left\lfloor \frac{1}{N} \sum_{i=1}^{N} OD_i \right\rfloor\right) \tag{25}$$

5.1.10 Correlação de grau

Como visto na Seção 4, uma das características que distinguem uma rede como sendo complexa é a lei de potências seguida por ela. Esta característica está associada diretamente com o conceito de distribuição de graus, o qual é o tema desta medida. Muitas vezes, é interessante verificar como os graus se distribuem sobre a rede, verificando-se a probabilidade de um nó possuir grau de saída e/ou entrada k_{out} e k_{in}, respectivamente. Outro ponto interessante de análise na rede é a correlação de graus dos vértices. Tal análise corresponde em observar se vértices interligados apresentam graus semelhantes.

Duas abordagens comuns são tomadas para cálculo e interpretação da correlação de grau. Uma primeira seria calcular a probabilidade $P(k_1,k_2)$, isto é, calcular a probabilidade de um vértice de grau k_1 estar ligado[22] com um vértice de grau k_2, ou alternativamente da probabilidade condicional $P(k_1 \big| k_2)$. A segunda abordagem, que é utilizada neste projeto, é baseada no coeficiente de Pearson, originalmente calculado como:

$$r = \frac{\frac{1}{M}\sum_{j>i} k_i k_j a_{ij} - \left[\frac{1}{M}\sum_{j>i} \frac{1}{2}(k_i + k_j) a_{ij}\right]^2}{\frac{1}{M}\sum_{j>i} \frac{1}{2}(k_i^2 + k_j^2) a_{ij} - \left[\frac{1}{M}\sum_{j>i} \frac{1}{2}(k_i + k_j) a_{ij}\right]^2} \tag{26}$$

na qual M representa o número de total arestas, k_i o grau do vértice i e a_{ij} um elemento da matriz de adjacência. A partir desta definição surge uma nomenclatura para a rede, dependendo do coeficiente de Pearson. Se r > 0 então a rede é chamada *assortativa*, caso contrário é chamada *não-assortativa*. Muitas vezes a assortatividade está relacionada com a estrutura topológica da rede. Como exemplo desta diferenciação, pode-se citar que redes sociais tendem a ser

[22] O termo *ligado* não necessariamente significa ligação entre arestas, em cálculos de correlação de grau, embora esta definição é aqui utilizada, por ser a abordagem mais comum.

assortativas, enquanto redes biológicas normalmente se apresentam não assortativas. Neste projeto, a correlação é calculada a partir da montagem de uma relação binária (x,y) cujo domínio e imagem são os graus e a relação segue o o sentido das arestas. Com isto, calcula-se o coeficiente de Pearson a partir de:

$$\mathrm{r} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\left[\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2\right]^{1/2}} \tag{27}$$

onde $\overline{x}$ representa a média dos pontos obtidos x e $\overline{y}$ representa a média dos pontos obtidos y. Neste caso r $\simeq 0$ representa que não existe nenhuma relação entre graus entre vértices ligados e r $\simeq \pm 1$ representa alta correlação entre vértices interligados.

5.1.11 Entropia de Entrada e Saída

Uma medida também relacionada com a distribuição de graus na rede e sua robustez é a Entropia, também conhecida como Entropia de distribuição dos graus. Trata-se de uma medida que quantifica a homogeneidade da rede. É definida como:

$$\mathcal{H} = -\sum_{k} P(k) \log P(k) \tag{28}$$

com P(k) representando a probabilidade de um vértice possuir grau k. Se a rede é homogênea, então P(k) $\simeq 1$, e o termo logarítmico tenderá a zero. Assim, quanto mais longe de zero (ou correspondentemente 1), mais heterogênea se comportará a rede. A Figura 6 mostra o comportamento da função entropia relativa à função $\left|\mathrm{xlog(x)}\right|$, o qual confirma a afirmação anterior. Define-se aqui dois tipos de entropia, entropia de entrada e saída correspondendo respectivamente ao cálculo de P(k) considerando grau de entrada e grau de saída da rede.

5.1.12 *Betweenness Centrality* e CPD

Medidas de centralidade, como *betweenness centrality*(B_v), buscam classificar a importância dos vértices ou arestas da rede. Para isso, o *betweenness centrality* classifica um dado vértice utilizando o número de caminhos mínimos que passam por este vértice. Mais precisamente, a definição de *betweenness centrality* para um dado vértice v é:

$$B_v = \sum_{i} \sum_{j} \frac{\sigma(i,u,j)}{\sigma(i,j)} \tag{29}$$

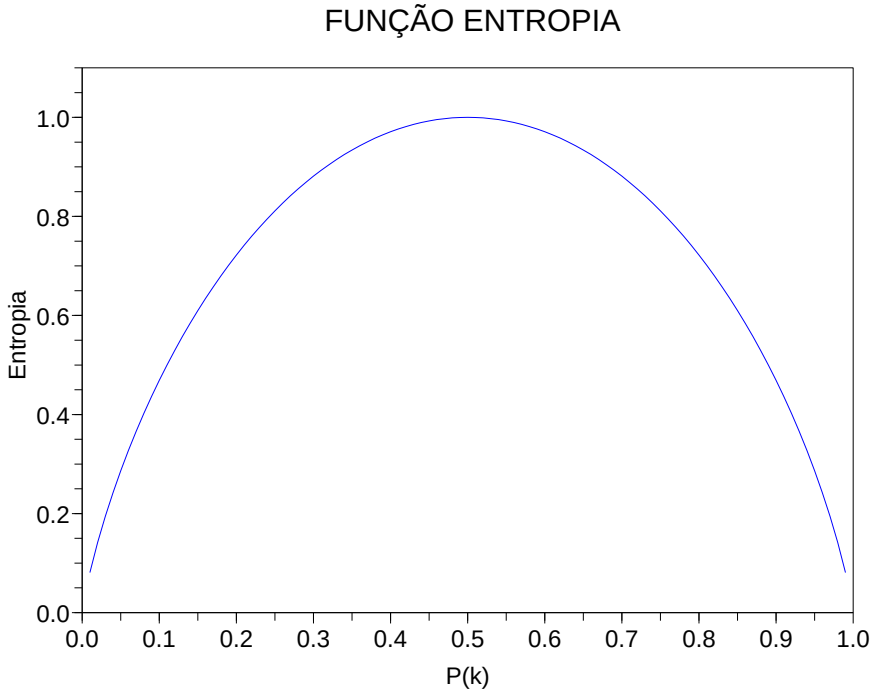

Figura 6: *Comportamento da função entropia.*

na qual o numerador da fórmula representa o número de caminhos mínimos que passam pelos vértices i, u e j e o denominador significa o número de caminhos mínimos que passam pelos vértices i e j. Outra medida diretamente derivável da medida *betweenness centrality* é a chamada *central point dominance* (CPD). Para isto, usa B_{max} como o máximo *betweenness* da rede e B_i como o *betweenness* do vértice i, como segue:

$$CPD = \frac{1}{N-1} \sum_i (B_{max} - B_i) \tag{30}$$

Para se ter uma idéia do significado desta medida, note que CPD será exatamente 1 quando o grafo apresentar uma topologia tipo estrela, isto é, existe um vértice central que contém todos os caminhos mínimos.

5.2 Medidas hierárquicas

As medidas hierárquicas podem ser entendidas como uma extensão das medidas padrões, a fim de permitir uma análise mais profunda da estrutura da rede. Também, com a definição de hierarquias, é possível definir métricas específicas a este nível de caracterização.

A fim de definir o conceito de hierarquias, considere o conjunto de vizinhos imediatos de um dado nó i. Denota-se tal conjunto por R_1(i). A importância deste conjunto está no fato de que

seus elementos são considerados como vértices intermediários na transmissão da informação do nó i a um dado nó j, $j \notin R_1(\mathrm{i})$. Portanto, pode-se dizer que tal conjunto é capaz de formar "arestas virtuais"(Costa, 2004) entre i e j. Matematicamente, define-se $R_1(\mathrm{i})$ como sendo o vetor $\overrightarrow{\nu}$, tal que :

$$\overrightarrow{\nu}(\mathbf{1}) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \end{pmatrix}$$
$$\overrightarrow{\nu}(\mathbf{2}) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \end{pmatrix}$$
$$\overrightarrow{\nu}(\mathbf{3}) = \begin{pmatrix} 0 & 0 & 1 & \cdots & 0 \end{pmatrix}$$
$$\overrightarrow{\nu}_{\mathbf{1}}(\mathbf{i}) = W(i,j) \cdot \overrightarrow{\nu}(i)$$

Agora, usando a função *delta de Kronecker* sobre o vetor $\overrightarrow{\nu}$, define-se $\overrightarrow{\rho_1}(i)$ como:

$$\overrightarrow{\rho_1}(i) = \delta(\overrightarrow{\nu}_1(i)) \tag{31}$$

fazendo com que $\rho_1(\mathrm{i})$ seja 1 quando $i \in R_1(\mathrm{i})$ somente, caso contrário, seu valor é nulo. De maneira análoga ao cálculo de $\overrightarrow{\nu}_1(\mathrm{i})$, pode-se estender seu conceito para $\overrightarrow{\nu}_d(\mathrm{i})$ como descrito na equação 32 :

$$\overrightarrow{\nu}_d(i) = W^d \cdot \overrightarrow{\nu}(i) \tag{32}$$

Assim, analogamente à equação 31, pode-se definir $\overrightarrow{\rho_d}(\mathrm{i})$ como o conjunto de vértices que estão a uma distância máxima de d arestas do vértice i, como descrito na equação 33.

$$\overrightarrow{\rho_d}(i) = \delta\Big(\sum_{k=1}^{d} \overrightarrow{\rho_k}(i) + \overrightarrow{\nu}(i)\Big) \tag{33}$$

Com estas definições, é possível conhecer o raio à distância i como na equação 34 :

$$\overrightarrow{\Gamma}_d(i) = \overrightarrow{\rho_d}(i) - \overrightarrow{\rho}_{d-1}(i) \tag{34}$$

ou simplesmente $R_d(\mathrm{i})$, segundo notação anterior. Agora, seja γ_d a sub-rede definida pelos nós em $R_d(\mathrm{i})$ mais as arestas entre os vértices de $R_d(\mathrm{i})$. Define-se nível hierárquico d da rede como γ_d mais as arestas para γ_{d+1} [23]. A seguir, algumas medidas intrínsecas à definição de hierarquias são explicitadas.

[23]Embora os passos descritos na definição de hierarquia ser mais claro, este método de cálculo não é o mais eficiente, sendo preferível utilizar outros métodos mais eficientes (Cormen et al., 2002) para redes maiores.

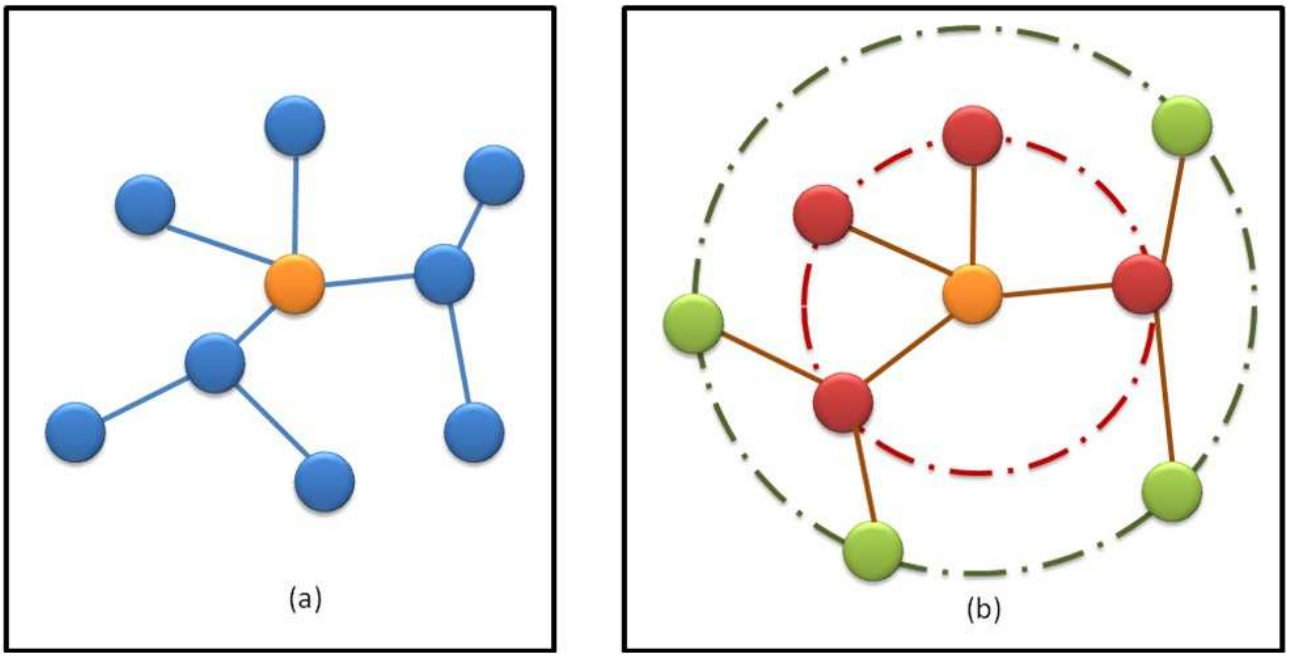

Figura 7: *Representação de uma rede com nó de referência em laranja no formato (a) não hierárquico e (b) hierárquico de dois níveis. Note que para o nível um, todos os nós contidos na circunferência vermelha serão representados por um único vértice. Para os outros níveis, a analogia é direta.*

5.2.1 Grau e *cluster coefficient* hierárquicos

A caracterização de graus e *cluster coefficients* para hierarquias basicamente estende o conceito de um vértice v para seu nível hierárquico i ao aderir ao vértice v seus vizinhos de distância máxima i, sendo que sua nova aresta será representada pelas arestas dos vértices de distância i que ligam a vértices de distância k, com $k \succ i$, conforme ilustra a Figura 7.

5.2.2 *Intra Ring Degree*

Para um dado vértice i no nível hierárquico d, esta medida é calculada como a média dos graus dos vértices na subrede definida por $\gamma_d(\mathrm{i})$.

5.2.3 *Inter Ring Degree*

Esta métrica refere-se ao número médio de conexões entre cada vértice nos anéis $R_d(\mathrm{i})$ e $R_{d+1}(\mathrm{i})$.

5.2.4 *Hierarchical Common Degree*

Esta métrica é calculada como sendo a média dos graus dos vértices em $R_d(\mathrm{i})$, considerando também as arestas da rede original. Desta forma, expressa a média do grau do vértice a cada

nível hierárquico, indicando como o grau da rede varia conforme a variação do nível hierárquico de análise.

6 Redes Complexas em PLN

Na medida em que várias áreas começaram a empregar as redes complexas para tarefas com resultados positivos, novas possibilidades foram se tornando mais claras. Uma destas é o uso de tal modelagem de rede para o Processamento de Língua Natural. Um dos clássicos experimentos evidenciando a característica de rede complexa é descrita em Cancho e Solé (2001) na qual foi montada uma rede derivada do *British National Corpus*, com os vértices representando as palavras, e as arestas conectando palavras que aparecem no córpus pelo menos uma vez, em seqüência ou separadas por uma palavra. Essa rede contém 478773 nós e 1,77 x 10^7 arestas. Também, outra rede foi construída, semelhante à anterior, com a diferença de que apenas são considerados os pares de palavras consecutivas (i,j) que ocorrem mais vezes do que seria esperado quando a independência entre as palavras é assumida, ou seja, quando $p_{ij} \succ p_i p_j$. Essa rede apresenta 460902 nós e 1,61 x 10^7 arestas. Foi mostrado que as duas redes apresentam as características *small-world* e *scale free*. Separadamente, Sigman e Cecchi (2002) também mostraram características que distinguem a rede de palavras como sendo uma rede complexa. A modelagem da *Wordnet* (Miller, 1985) em uma rede, no qual os nós são os substantivos e as arestas são as relações semânticas entre vértices (como antonímia e hipernímia) mostrou que a polissemia afeta a organização da rede, novamente caracterizando-a como uma rede *small-world*. Em uma outra abordagem Motter et al. (2002) modelaram um *thesaurus* da língua inglesa conectando duas palavras relativamente próximas semanticamente, o que permitiu averiguar-se a propriedade *small-world* e *scale free*. Assim, ficou claro que várias redes lingüísticas poderiam ser modeladas com sucesso como redes complexas.

O uso de redes na área de PLN não é algo recente. Exemplos de tais aplicações foram aplicadas por exemplo em sumarização (Mihalcea, 2006), desambigüização de sentido, e análise de sentimento (Pang e Lee, 2004), entre outros. No entanto, apenas recentemente as redes complexas foram o enfoque da análise da linguagem, caracterizado pelo uso de técnicas da física estatística aliada aos estudos lingüísticos (Dorogovtsev e Mendes, 2001). Pode-se dizer que as principais pesquisas concentram na modelagem dos textos como redes de adjacência a fim de realizar algum processamento sobre tal modelagem, sendo vários os exemplos relacionados. Por exemplo, a escolha do sinônimo mais provável para um contexto é feito a partir de uma rede complexa em Edmonds (1997). A avaliação de qualidade a partir da correlação de métricas extraídas da rede e a definição de indicadores de qualidade discretos e contínuos foram propostos com bons resultados para distinguir redações em português de estudantes brasileiros no exame nacional do ensino médio (ENEM) (Antiqueira et al., 2007). Ainda quanto à quali-

dade, é importante citar os trabalhos de avaliação em qualidade de sumários (Antiqueira et al., 2009), (Pardo et al., 2006) e traduções automáticas (Amancio et al., 2008). Particularmente, nesta última, buscou-se fazer a correlação entre medidas da rede fonte e alvo através do mapeamento por alinhamento lexical[24] (Veronis, 2000), o qual permitiu expressiva distinção entre tradutores. Por fim, pode-se citar os trabalhos em caracterização de autoria por métricas (Antiqueira et al., 2006), estabilização de métricas em redes lingüísticas (Margarido et al., 2008) e modelagens alternativas considerando sentenças inteiras como vértices (Caldeira et al., 2006) também como exemplos das aplicações recentes das redes complexas conjuntamente com a área de Processamento de Língua Natural.

A adoção de redes complexas evidencia uma das tendências dos sistemas de Processamento de Língua Natural (PLN): a alternância do formalismo puramente baseado no conhecimento armazenado nas gramáticas e em léxicos para introdução de formalismos mais abrangentes e robustos, a fim de se obter o máximo possível de entendimento do texto.

Enfim, pode-se o dizer que atualmente a associação entre redes complexas e aplicações de processamento de língua natural ainda é um campo amplo de pesquisa, uma vez que esta ainda não foi explorada ao ponto de se conhecer bem os limites e capacidades das redes enquanto representação de textos.

7 Metodologia : pré-processamento e construção das redes

A fim de se obter as características referentes a cada texto, removem-se as *stopwords*[25]. Além disso, as palavras restantes são lematizadas, a fim de agrupar conceitos de mesma forma canônica, mas com flexões diferentes. Adicionalmente, o texto é etiquetado morfossintaticamente, neste caso pelo *tagger* MXPost (Aires et al., 2000), baseado no modelo de Ratnaparki (1996), o qual adiciona informações úteis na resolução de ambigüidades na fase de lematização. Esta, por sua vez, é feita acessando-se o léxico computacional do NILC[26] (Nunes et al., 1996), no qual cada palavra tem uma regra associada para geração da forma canônica. Por exemplo, a forma pré-processada para a sentença *"Um substituto para a insulina injetável, medicamento para diabéticos, está em testes no Canadá"* em português e inglês é ilustrada na Tabela 2.

A estrutura que representa a rede derivada de um texto é uma matriz de adjacências com pesos. Após o pré-processamento do texto, as N palavras distintas restantes passam a representar os nós da rede, e a seqüência de palavras resultante é utilizada na criação das arestas,

[24] O alinhamento lexical produz indicações de quais palavras (ou unidades multipalavra) fonte estão alinhadas com quais palavras (ou unidades multipalavra) alvo.

[25] *Stopwords* são palavras de pouco significado semântico, tal como artigos, preposições e outros.

[26] O léxico do NILC (Núcleo Interinstitucional de Lingüística Computacional), conta com cerca de 1,5 milhão de palavras. É o maior léxico computacional para o português brasileiro.

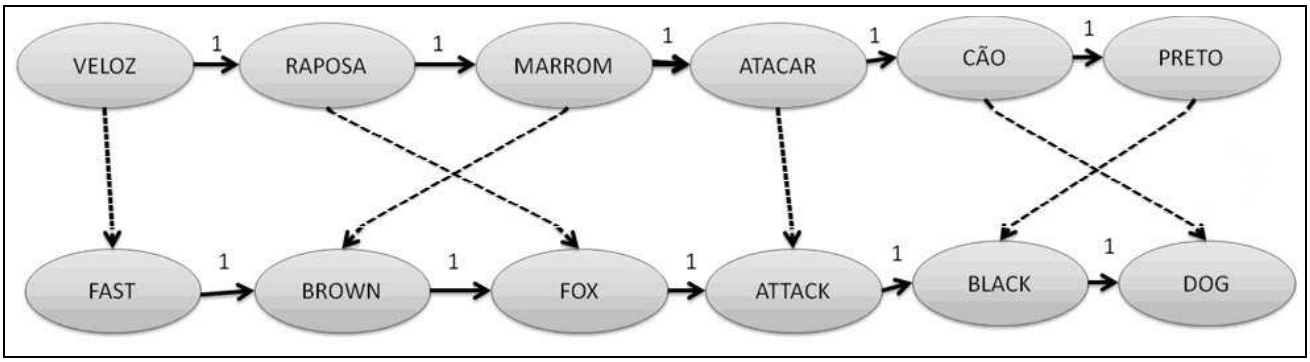

Figura 8: *Exemplo de rede formada para a sentença "A veloz raposa marrom ataca o cão preto". As setas espessas representam as arestas da rede fonte e da rede alvo e as setas tracejadas representam a indicação de qual palavra foi traduzida em qual palavra.*

de modo que, para cada par de palavras consecutivas, existe uma aresta direcionada correspondente na rede. As arestas também apresentam pesos, os quais indicam o número de vezes que as respectivas associações de palavras aparecem no texto. Todas as métricas são obtidas da matriz de adjacências W que representa a rede. Ela apresenta inicialmente todos os elementos iguais a zero e, a cada par de palavras (i,j) lido do texto, faz-se $W_{ji} = W_{ji} + 1$, incrementando desse modo o peso da associação i → j. A Figura 8 ilustra uma rede construída para a sentença *"A veloz raposa marrom ataca o cão preto"*.

Texto original	**Texto pré-processado**
Um substituto para a insulina injetável, medicamento para diabéticos, está em testes no Canadá.	substituto insulina injetável medicamento diabético estar teste Canadá.
A substitute for injectable insulin, a medicine for diabetics, is under test in Canada.	substitute injectable insulin medicine diabetic test Canada.

Tabela 2: *Exemplo de pré-processamento de uma sentença em português e de uma senteça em inglês.*

8 Experimentos

8.1 Estudo do efeito de nova modelagem sobre métricas

Este experimento tem como objetivo avaliar o efeito da modelagem descrita em Cancho e Solé (2001) sobre as medidas das redes complexas, que difere da modelagem descrita na Seção 7 já

que esta última conectou por arestas apenas as palavras imediatamente adjacentes, de forma a gerar uma conexão conceitual entre palavras vizinhas. Esta ligação semântica entre palavras adjacentes não permite obviamente uma ligação direta entre expressões do tipo *trem de ferro vermelho*. Assim, a princípio, conceitos semelhantes relacionados ao *trem* possuem níveis de conectividade distintas, já que enquanto *ferro* está diretamente relacionado a *trem*, *vermelho* não se relaciona com *trem*, ou melhor, a relação é apenas indireta. Para permitir também a conexão direta, este experimento modelou as redes de forma que palavras de distância igual a 2 também pudessem ser ligadas diretamente.

A partir da nova modelagem, medidas globais foram extraídas de um corpus de 100 textos em inglês e comparadas com as mesmas medidas dos mesmos textos da modelagem tradicional (apresentada na Seção 7). Em seguida, medidas globais de grau de entrada (apresenta o mesmo valor que o grau de saída), coeficiente de aglomeração, SP_1, SP_2 e SP_3 foram extraídas. A Figura 9 ilustra a diferença percentual entre a medida obtida com apenas arestas simples (AS) e com arestas simples e arestas de distância igual a 2 (AD). Nesta figura fica evidenciado que praticamente em todos os 100 textos a medida de coeficiente de aglomeração apresentou um aumento de -100% de AS em relação AD. Semelhantemente, a medida de grau apresentou uma variância ainda mais apertada, refletindo um aumento de -50% de AS em relação a AD. O primeiro resultado é intuitivo, já que a adição de arestas na rede muito provavelmente faz a aglomeração da rede aumentar, pela própria definição da medida. Também o resultado para graus é muito intuitivo, uma vez que ao adicionar arestas tipo AD, praticamente o número de arestas totais na rede é dobrado, refletindo portanto a pouca variância obtida em torno do ponto de 50%. Os caminhos mínimos seguem raciocínio análogo, mas agora com um aumento positivo entre AS e AD, já que arestas adicionais fazem encurtar os caminhos. É interessante observar que os valores de caminhos mínimos de AD parecem cair pela metade, uma vez que agora por onde se percorria através das AS, pode-se percorrer agora pelas AD. Este fato é refletido pela média se aproximar de 50% nos três tipos de caminhos mínimos.

Um fato interessante destes experimentos foi que as medidas de grau de entrada, grau de saída e coeficiente de aglomeração globais parecem apenas ter sido multiplicadas por uma constante com a adição destas novas arestas. Isto porque a variância da diferença relativa apresentou-se muito pequena, o que sugere provavelmente a mesma multiplicação por constante em nível local, isto é, em nível nodal. Desta forma, pode-se supor que medidas locais análogas também sejam multiplicadas por uma constante. Assim, por exemplo, provavelmente os coeficientes angulares e de Pearson (dependentes de métricas locais), de grande importância na distinção de traduções no trabalho realizado em Amancio et al. (2008), pouco mudarão com esta abordagem, uma vez que constantes são canceladas (considerando-se o fato da pequena variância obtida). Portanto, como conclusão final deste experimento, pode-se admitir que praticamente esta modelagem não afeta o resultado dos experimentos para análise de textos.

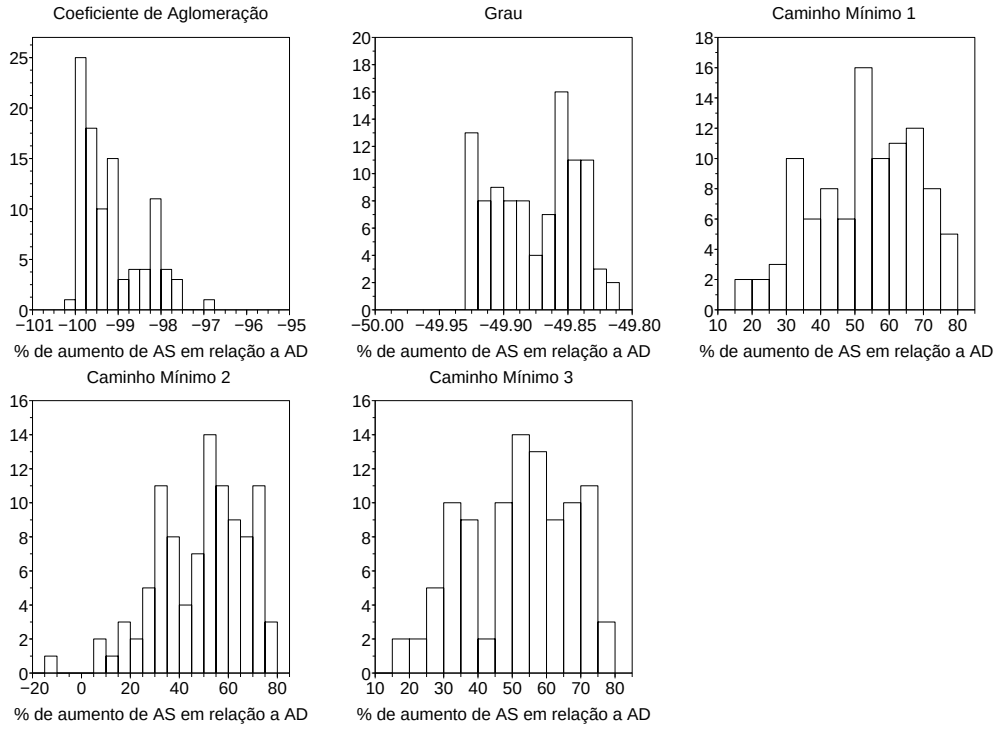

Figura 9: *% relativa de aumento entre a medida obtida pelo modelagem tradicional deste trabalho (AS) e a modelagem com adição de arestas de distância 2 (AD).*

8.2 Tradução automática

8.2.1 Estabilização de métricas

Uma das maneiras de avaliar um tradutor automático, especificamente aqueles bidirecionais, é comparar várias traduções bidirecionais no mesmo idioma, por exemplo, o idioma do texto fonte. Assim, intuitivamente poderia-se imaginar que os tradutores de boa qualidade tendem a não variar as traduções obtidas consecutivamente, ou melhor, tendem a variar de forma mínima. Um exemplo de uma tradução incorreta, gerada pelo tradutor do Google, com língua intermediária chinês simplificado é dado abaixo:

```
A veloz raposa marrom ataca o cão preto
-> N traduções ->
Uma rápida raposa marrom ataca o cão preto
```

Mais especificamente, o processo de traduções realizado no experimento é ilustrado na Figura 10. Inicialmente, traduziu-se o texto originalmente em inglês para o português, gerando o texto intermediário, o qual por sua vez foi traduzido do português para o inglês. Este processo foi

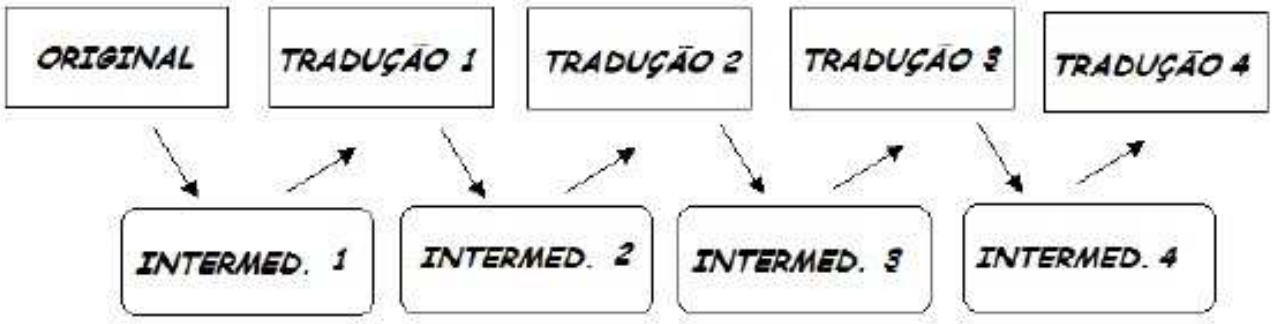

Figura 10: *Sequência de passos no experimento com várias traduções do mesmo texto. Textos originais e tradução pertencem ao inglês e os textos intermediários pertencem ao português.*

repetido cinco vezes para cada texto no corpus (10 no total, retirados da seção Laboratório da Revista Pesquisa Fapesp), para cada tradutor automático (Free Translation e Intertran). Obtidas as traduções automáticas, retirou-se as medidas globais de coeficiente de aglomeração, grau[27] e *CDD*. Observando a variação das medidas, pretende-se verificar sua estabilidade para cada tradutor automático, possivelmente relacionando a presença ou ausência de estabilidade da medida com a qualidade da tradução.

A Figura 11 ilustra o comportamento das traduções sucessivas quanto à medida de coeficiente de aglomeração. Observa-se neste caso que em mais da metade dos textos, tal medida parece se estabilizar para ambos os textos. Da mesma forma, praticamente não existe nenhum texto na qual nenhum dos tradutores automáticos deixou de estabilizar tal medida ao final das quatro traduções. Este mesmo comportamento e distribuição foi encontrado quando a dinâmica no desvio das componentes (*CDD*) foi analisada. Também pôde-se observar uma leve tendência, quando analisada estas duas medidas, de apenas a tradução do Intertran se estabilizar, para um dado texto. A Tabela 3, segundo a porcentagem de textos que tiveram as medidas estabilizadas pelos tradutores (individualmente, ambos ou nehum), evidencia estas conclusões.

Já a Tabela 4 apresenta a porcentagem de textos que estabilizaram sua medida no intervalo de 4 traduções, mas agora não por métrica, mas sim por tradução. Assim, fica claro observar que o melhor tradutor se aproxima do pior apenas na medida de grau, medida na qual o pior tradutor deixa de apresentar o comportamente de estabilidade. Contrariamente, o melhor tradutor apresenta aproximadamente a mesma taxa de estabilidade (um pouco maior de 50%) independentemente da medida analisada.

O padrão a ser seguido neste corpus, portanto, mostra que a tradução boa tende a manter a taxa de textos que se estabilizam em 4 traduções sucessivas, taxa na qual se apresenta pouco maior que 50%. Do outro lado, o tradutor ruim agrega a estabilidade à medida analisada. Ao considerar todas as medidas com a mesma ponderação na importância de estabilidade, também

[27]Para medidas globais o média de entrada é o mesmo que o grau médio de saída.

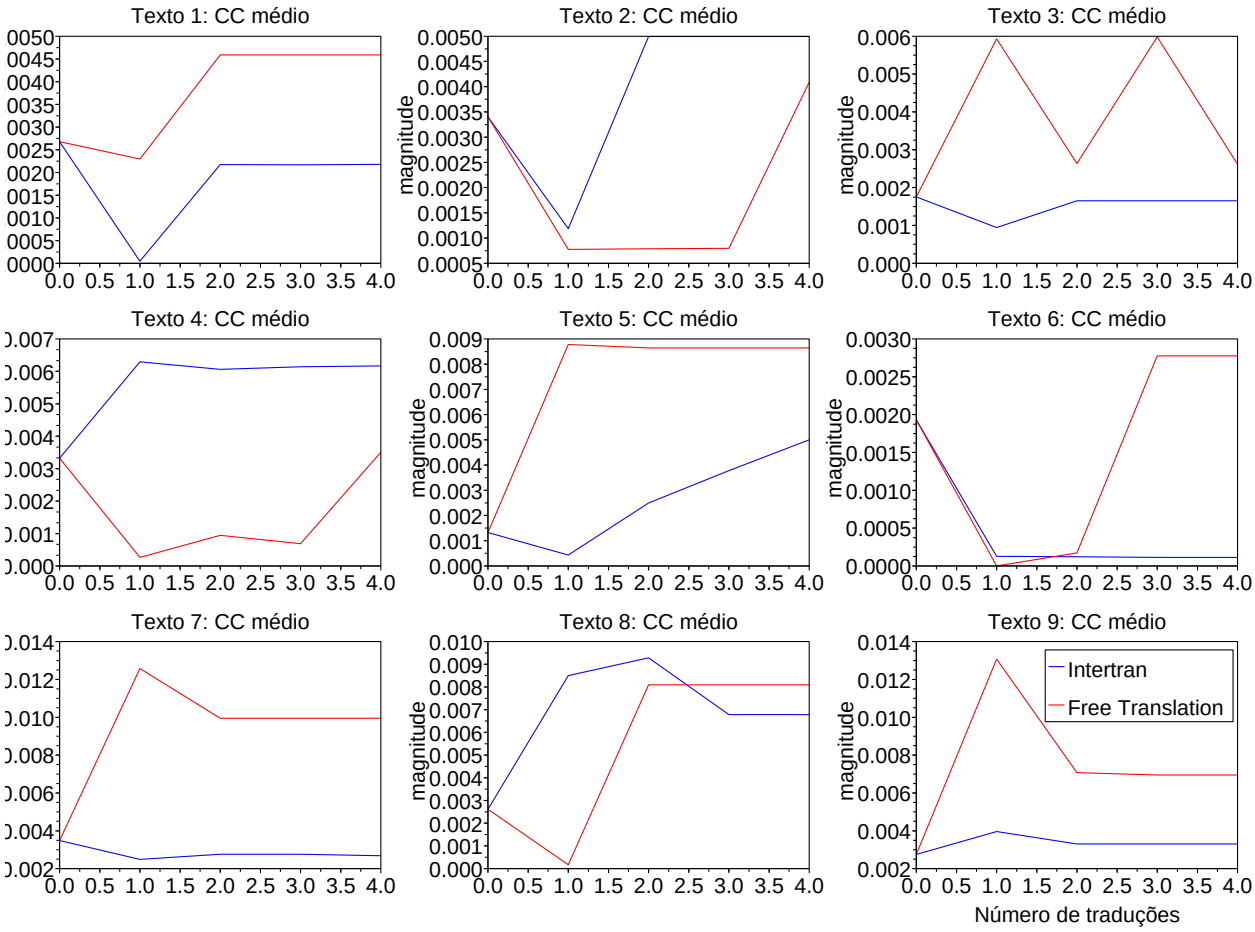

Figura 11: *Sequência de passos no experimento com várias traduções do mesmo texto. Textos originais e tradução pertencem ao inglês e os textos intermediários pertencem ao português.*

Tradução	CDD	OD	CC
Intertran apenas	22%	11%	33%
Free Trans. apenas	11%	11%	11%
Intertran e Free	67%	45%	55%
Nenhum	0%	33%	0%

Tabela 3: *Tabela da porcentagem de textos cujas medidas foram estabilizadas, segundo cada métrica.*

Tradução	CDD	OD	CC	% TOTAL
Intertran	89%	44%	89%	**74%**
Free Translation	56%	56%	67%	**59%**

Tabela 4: *Tabela da porcentagem de textos cuja medida foram estabilizadas, segundo cada tradução.*

pode-se concluir que o tradutor ruim apresenta uma maior tendência de se estabilizar em relação ao melhor tradutor. Por fim, pode-se perceber que o número de traduções necessárias para que as medidas se estabilizem provavelmente não é muito maior que quatro, segundo as porcentagens extraídas da última coluna da Tabela 4.

8.2.2 Distinção por níveis hierárquicos

Este experimento tem como objetivo reconhecer certos padrões presentes na dinâmica das redes complexas, espeficicamente na sua estrutura definida pelas medidas hierárquicas. Os experimentos referentes foram realizados focando a distinção de traduções automáticas comprovadamente de qualidade opostas (Oliveira Jr. et al., 2000), sendo que as línguas envolvidas foram o português, o espanhol e o inglês. Todos os tradutores automáticos utilizados são de uso gratuito e disponíveis por interface web. Estes são: Apertium[28] (para traduções envolvendo o espanhol e o português), Google[29] (para traduções envolvendo o inglês e o português) e Intertran[30] (para traduções envolvendo o espanhol, português e o inglês). Em todos os casos, o português foi tomado sempre como língua alvo (língua de saída do tradutor), sendo que o espanhol e o inglês alternaram-se nos experimentos como língua fonte (língua de entrada no tradutor).

O procedimento padrão inicial para avaliação de qualidade de tradução do texto fonte (TF) para o texto alvo (TA) consistiu em extrair as métricas descritas na Seção 5.2 para cada vértice v_f da rede fonte e para cada vértice v_a da rede alvo, variando os níveis hierárquicos de 1 até N, sendo N = $\min\{v_a, v_f\}$. Com isto, tanto cada vértice do TF como cada vértice do TA acabam sendo caracterizados por duas matrizes MF e MA respectivamente :

[28] $http://xixona.dlsi.ua.es/apertium$
[29] $http://www.google.com/language\_tools?hl=en$
[30] $http://www.tranexp.com:2000/Translate/result.shtml$

$$\mathbf{MF_{vert}} = \begin{pmatrix} \mu_{f1}^{A} & \mu_{f1}^{B} & \mu_{f1}^{C} & \cdots & \mu_{f1}^{X} \\ \mu_{f2}^{A} & \mu_{f2}^{B} & \mu_{f2}^{C} & \cdots & \mu_{f2}^{X} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_{fN}^{A} & \mu_{fN}^{B} & \mu_{fN}^{C} & \cdots & \mu_{fN}^{X} \end{pmatrix}$$

$$\mathbf{MA_{vert}} = \begin{pmatrix} \mu_{a1}^{A} & \mu_{a1}^{B} & \mu_{a1}^{C} & \cdots & \mu_{a1}^{X} \\ \mu_{a2}^{A} & \mu_{a2}^{B} & \mu_{a2}^{C} & \cdots & \mu_{a2}^{X} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_{aN}^{A} & \mu_{aN}^{B} & \mu_{aN}^{C} & \cdots & \mu_{aN}^{X} \end{pmatrix}$$

Enfim, $MY_{vert} = \{u_{ij}\}$, na qual Y representa a rede fonte (F) ou a rede alvo (A), i representa o nível hierárquico i e j representa a respectiva métrica de análise para o vértice em questão. Dado um vértice i da rede fonte, define-se a relação binária F : i $\rightarrow$ j, com j pertencente ao conjunto de vértices da rede fonte da seguinte forma :

$$F_h(i) = j | j \in redealvo \tag{35}$$

e $Erro_h$(i, j), definido por 36, é o mínimo erro entre todos os j :

$$\epsilon_h(i,j) = \sum_{k=1}^{N} \Big(MF_i(h,k) - MA_j(h,k) \Big)^2 \tag{36}$$

Adicionalmente, a fim de se evitar que métricas de valores absolutos superiores possam ter maior ponderação no cálculo da função erro, cada parcela da soma é normalizada.

Com a definição do $Erro_h(i,j)$ e consequentemente da função F(i), define-se a taxa de acerto T_i para o vértice i da rede fonte. Seja A o conjunto de vértices obtidos da relação F_h(i) e B o conjunto de vértices obtido do alinhamento lexical L(i) realizado pelo LIHLA [31] (Caseli e Nunes, 2005). Assim, para um dado vértice i, sua taxa de acerto no nível h é dada por:

$$T_h(i) = \frac{|A \cap B|}{|B|} \tag{37}$$

A partir desta definição, segue que a taxa de acerto do par TF-TA será dada por :

$$T_{af} = \frac{1}{N} \sum_{i=1}^{N} T_h(i) \tag{38}$$

A motivação do uso desta taxa de acerto provém das altas taxas do experimento realizado

[31] Em uma tradução de n_s palavras numa sentença fonte para uma sentença alvo com n_t palavras, o LIHLA tenta criar um alinhamento entre palavras desta sentença. Quando um alinhamento simples considera mais do que uma palavra na sentença fonte ou alvo, tais palavras são agrupadas em um único vértice da rede.

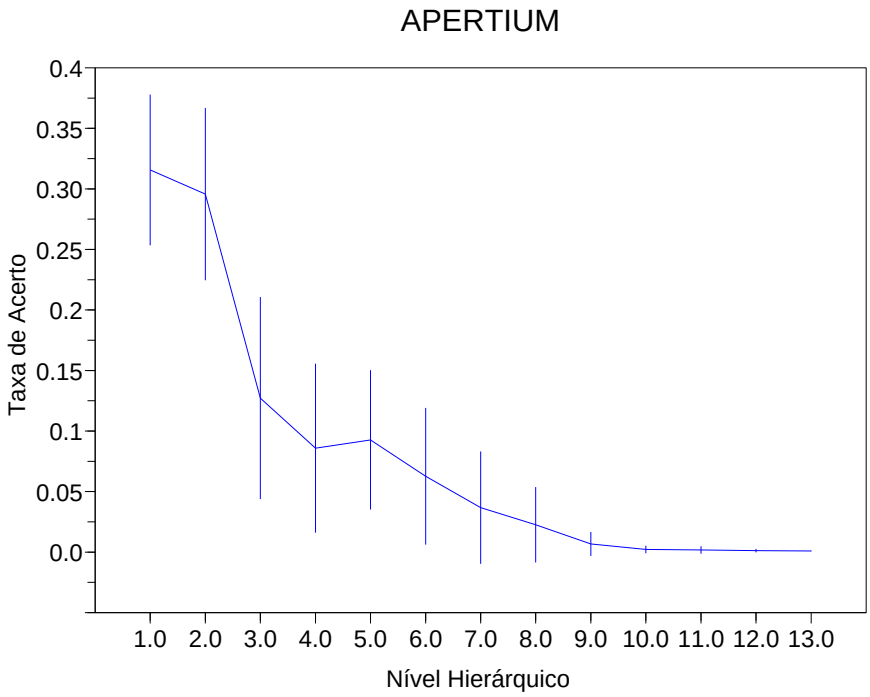

Figura 12: *Variação da taxa de acerto (com indicação do desvio padrão)segundo o nível hierárquico utilizado, para o tradutor automático de boa qualidade Apertium.*

em Costa e Rocha (2006), cuja rede fonte representa a modelagem da proteína *S. Cerevisiae* (Sprinzak e Margalit, 2001) e a rede alvo representa a rede formada pelos vértices da rede fonte cujas arestas foram rearranjadas aleatoriamente. A utilização de métricas hierárquicas se justificam pela sua importância quanto a sua variação e dinâmica ao longo dos níveis e por isso foi o objetivo de estudo nos experimentos que seguem.

Num primeiro experimento utilizou-se um corpus composto de 50 textos distintos provenientes de traduções automáticas (Apertium, Intertran) do espanhol para português, sendo cada rede formada aproximadamente de 300 vértices. O comportamento obtido para as taxas de acerto em função do nível hierárquico estão ilustrados nas figuras 12, 13 e 14 .

Pode-se concluir que para redes pequenas, a principal informação se concentra nos níveis hierárquicos inferiores, sendo que a taxa média de acerto e o desvio padrão diminui gradativamente com o avanço da análise em níveis maiores, como é ilustrada na Figura 12 para o tradutor Apertium e na Figura 13 para o tradutor Intertran. Isto pode ser explicado pelo fato de que muitos vértices na rede se tornam periféricos quando analisados em níveis hierárquicos de magnitude próximos ao comprimento da rede[32], o que faz com que sua capacidade de distinção seja reduzida. Também é interessante notar também o padrão exibido na Figura 14, o qual revela a taxa de acerto dos dois tradutores automáticos para o espanhol de qualidades distintas. Em

[32]Refere-se aqui ao maior caminho mínimo entre os caminhos mínimos de quaisquer dois vértices na rede.

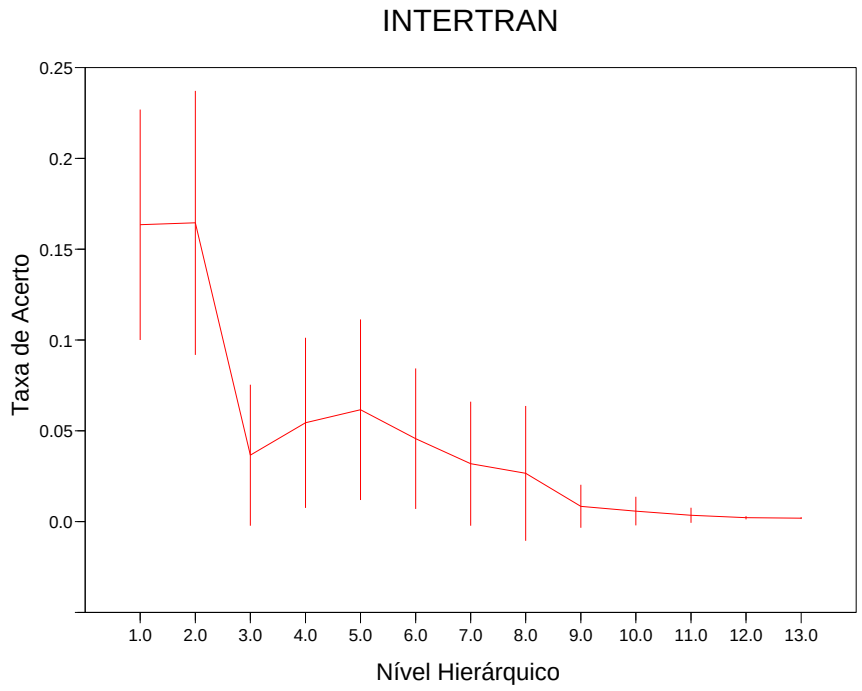

Figura 13: *Variação da taxa de acerto (com indicação do desvio padrão)segundo o nível hierárquico utilizado, para o tradutor automático de baixa qualidade Intertran.*

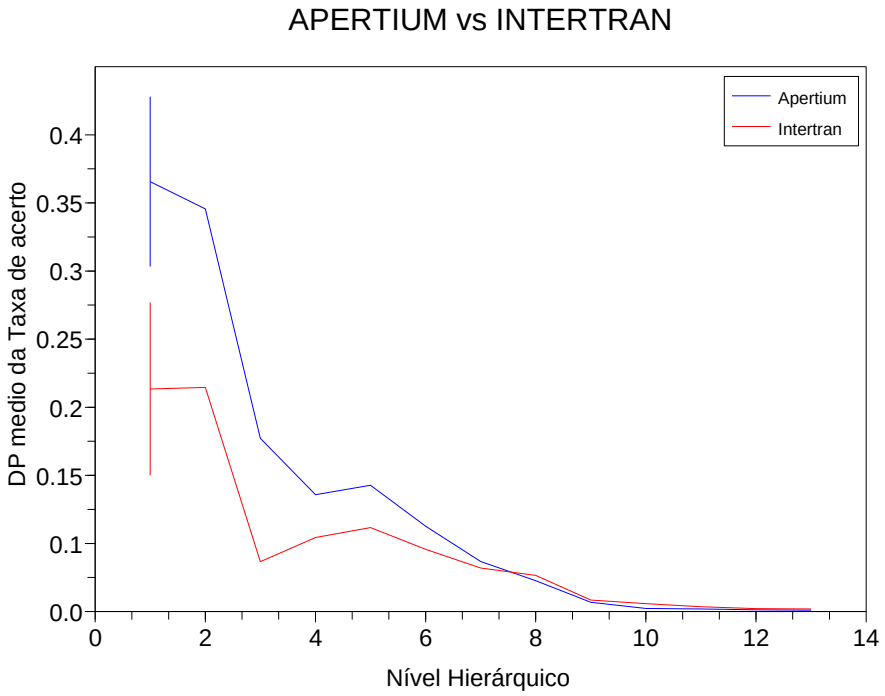

Figura 14: *Variação da taxa de acerto segundo o nível hierárquico utilizado para os tradutores automáticos do espanhol para o português Intertran e Apertium.*

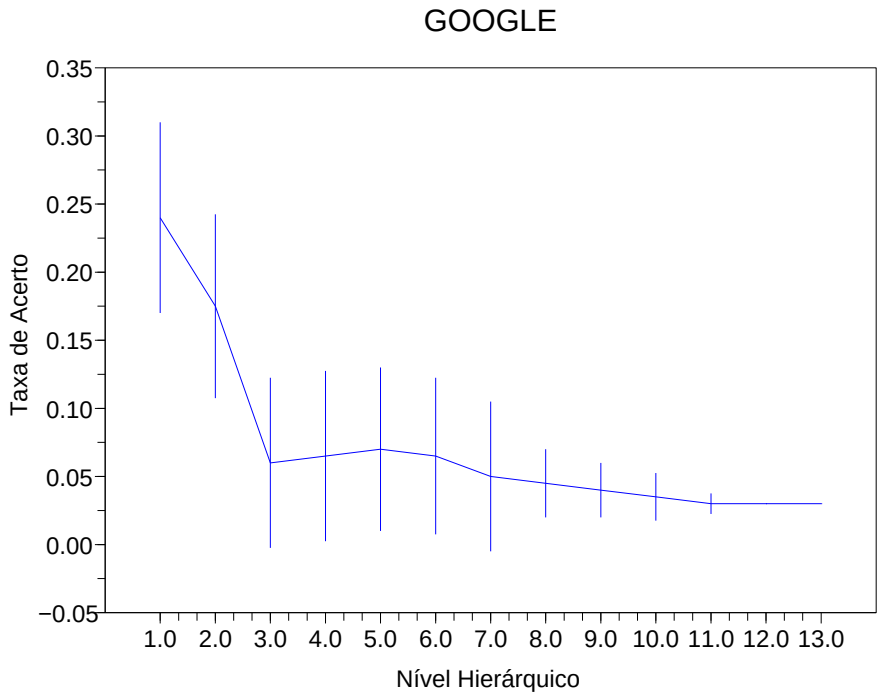

Figura 15: *Variação da taxa de acerto segundo o nível hierárquico utilizado para o tradutor automático Google.*

praticamente todos os níveis analisados, a taxa de acerto para o tradutor de alta qualidade foi superior à taxa de acerto do Intertran, evidenciando um possível padrão de distinção entre estes dois tipos de tradutores. Enfim, pode-se notar que a distinção é suficientemente boa entre tais tradutores.

Analogamente ao experimento anterior, projetou-se o experimento para o inglês, mas agora com 50 textos de aproximadamente 500 vértices cada, para os tradutores automáticos Google e Intertran. As curvas obtidas são descritas nas figuras 15, 16 e 17.

Como no caso para o espanhol, a taxa de acerto segue um comportamento padrão para ambas as traduções, caracterizado por uma queda ao longo dos níveis hierárquicos. No entanto, a dependência lingüística e a proximidade dos tradutores se torna mais complexa, refletindo uma difícil distinção a partir da taxa de acerto unicamente. Felizmente, uma razoável separação pode ser encontrada não na análise da taxa média de acerto, mas sim pela análise do desvio padrão da taxa de acerto dos textos para cada nível, como ilustra a Figura 17.

Em uma segunda análise, repetiu-se os experimentos citados anteriores com uma pequena mudança no cálculo da taxa de acerto. Utilizou-se para o cálculo da taxa de acerto no nível h informação das métricas não apenas do nível h, mas sim de todos os seus níveis inferiores, como denota a nova função erro definida pela equação 39. Analogamente a uma rede social, pode-se dizer que a identificação de um indivíduo não é feito apenas por si próprio, mas com ajuda de

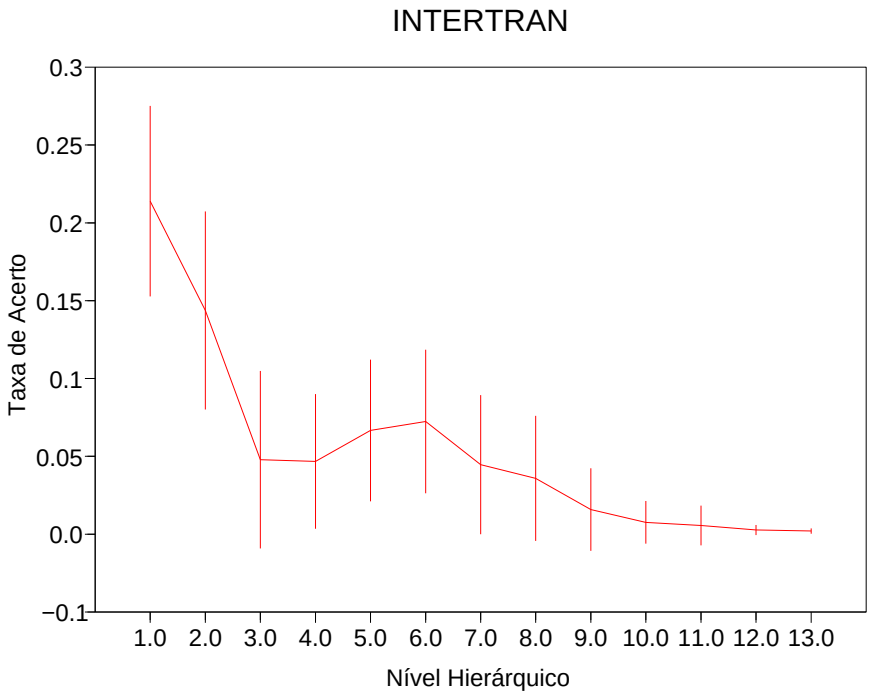

Figura 16: *Variação da taxa de acerto segundo o nível hierárquico utilizado para o tradutor automático Intertran.*

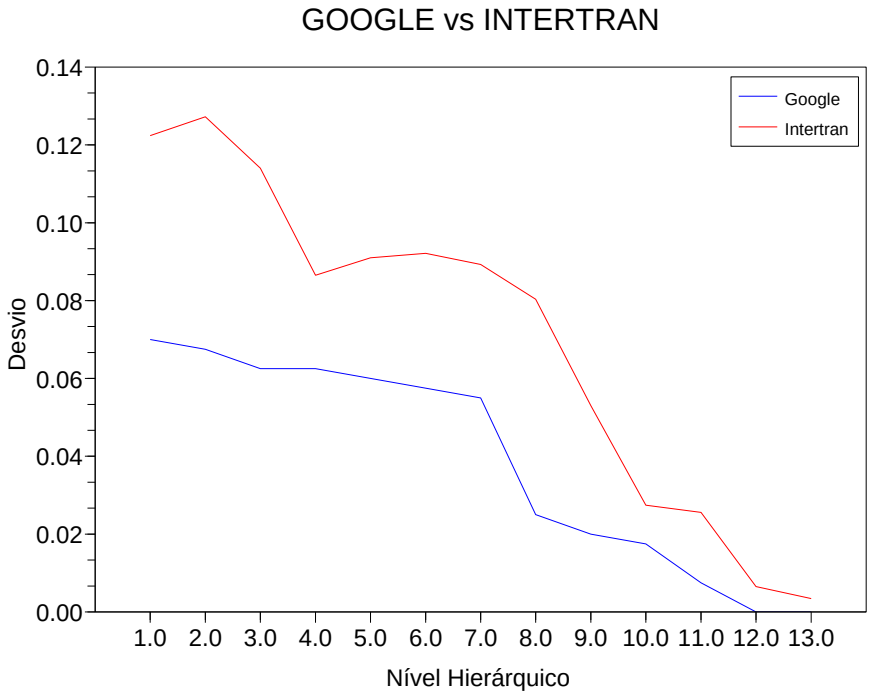

Figura 17: *Variação da taxa de acerto segundo o nível hierárquico utilizado para os tradutores automáticos do inglês para o português Intertran e Google.*

Algoritmo	Espanhol-Português	Inglês-Português
Regra	99%	87%
Árvore	98%	84%

Tabela 5: *Taxa de acerto na determinação do tipo de tradução para os pares espanhol-português (Apertium ou Intertran) e inglês-português (Google and Intertran) obtido com aprendizado de máquina usando mapeamento vértice a vértice através de níveis hierárquicos.*

parentes de tal indíviduo. Desta forma, diferentemente do caso anterior esperava-se haver um crescimento da taxa de acerto, devido ao aumento de informação não redundante utilizada.

$$\epsilon_H^{AC}(i,j) = \sum_{h=1}^{H}\sum_{k=1}^{N}\Big(MF_i(h,k) - MA_j(h,k)\Big)^2 \tag{39}$$

A Figura 18 ilustra o comportamento encontrado para ambos os tradutores (idioma espanhol) ao variar o nível hierárquico. Conforme suposto anteriormente, a curva de taxa de acerto obteve um aumento logo nos primeiros níveis hierárquicos, caracterizando um ganho de informação ao considerar uma profundidade maior na rede. Em seguida, houve uma estabilização e por fim uma queda, que pode ser explicada pela degeneração da rede em níveis hierárquicos profundos, o qual faz com que seja introduzido ruído no conjunto de métricas utilizadas.

Independentemente do comportamento das curvas, pode-se dizer que o objetivo esperado foi alcançado dado que um claro padrão de diferenciação de tradutores de qualidades opostas foi encontrado, como pode ser visto na Figura 19. Em todos os níveis hierárquicos estudados, a taxa de acerto do tradutor de melhor qualidade (Apertium) é superior à taxa de acerto do tradutor de qualidade inferior. Até mesmo considerando o desvio padrão contabilizado sobre e sob a média, em nenhum caso há uma intersecção no gráfico. De maneira análoga, os tradutores para o inglês seguem a dinâmica anterior, com melhor distinção pelo desvio padrão, como pode ser observado na Figura 20 e 21.

Adicionalmente, um estudo foi conduzido a fim de quantificar a eficiência deste método em distinguir a qualidade da tradução por meio da predição do tradutor utilizado. Para isto, utilizou-se aprendizado de máquina (Witten e Frank, 2005), com algoritmos de regra (Cohen, 1995), (Furnkranz e Widmer, 1994) e árvore de decisão (Quinlan, 1993), sendo que os atributos foram a taxa de acerto médio e desvio padrão em cada nível e como classificação o tipo de tradutor, dependente do par de idioma utilizado. Os resultados ilustrados na Tabela 5 refletem as considerações anteriores ao mostrar as taxas de acerto do classificador numa análise *10 fold cross validation* (Kohavi, 1995), com boa distinção entre tradutores automáticos especialmente no caso do par de idiomas português-espanhol.

Enfim, a caracterização de textos como redes complexas com análise de métricas hierárquicas mostrou-se como mais uma alternativa à avaliação da qualidade de traduções automáticas.

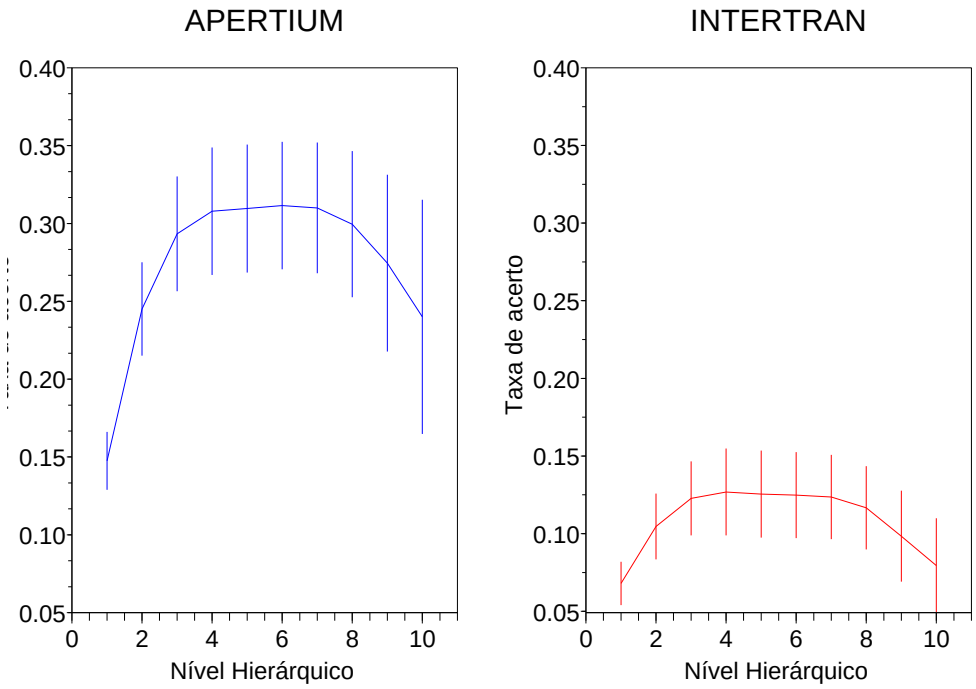

Figura 18: *Variação da taxa de acerto segundo o nível hierárquico utilizado para os tradutores automáticos Apertium e Intertran, utilizando a definição de erro segundo a equação 39.*

Desta forma, tal análise, juntamente com as técnicas já desenvolvidas em Amancio et al. (2008) podem evidenciar um rico conjunto de ferramentas alternativas aos métodos padrões de avaliação automática de traduções.

8.3 Simplificação Textual

8.3.1 Reconhecimento de padrões segundo o tipo de simplificação para métricas globais

Para este conjunto de experimentos de simplificação textual pretendeu-se encontrar padrões a partir das redes, que pudessem especificar qualitativa e/ou quantitativamente o grau de simplificação em que um dado texto se encontra. Desta forma utilizou-se as ferramentas e um subconjunto do corpus[33] disponibilizados no projeto PorSimples[34] , com as avaliações dos textos sendo obtidas com as métricas usuais de redes complexas. Especificamente, métricas foram extraídas e comparadas para um corpus composto de 125 textos obtidos de uma simplificação natural[35] e de uma simplificação do tipo forte[36]. Exemplos de tais simplificações são ilustrados

[33]Disponível em *http : //caravelas.icmc.usp.br/portal/index.php/texts/index*
[34]Simplificação Textual do Português para Inclusão e Acessibilidade Digital.
[35]Baixo nível de simplificação.
[36]Alto nível de simplificação.

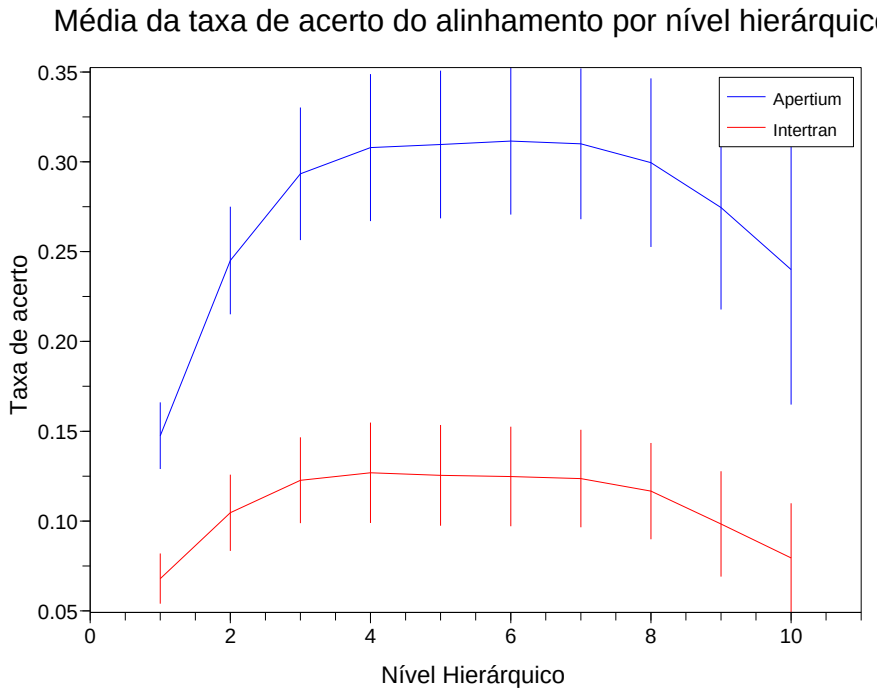

Figura 19: *Variação da taxa de acerto segundo o nível hierárquico utilizado para os tradutores automáticos Apertium e Intertran, utilizando a definição de erro segundo a equação 39.*

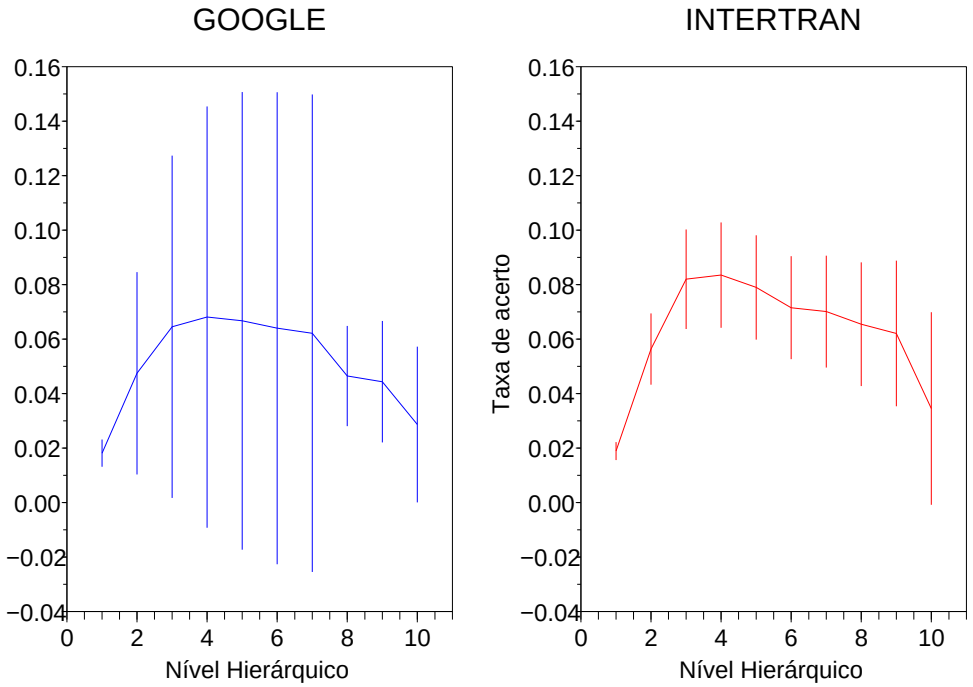

Figura 20: *Variação da taxa de acerto segundo o nível hierárquico utilizado para os tradutores automáticos Google e Intertran, utilizando a definição de erro segundo a equação 39.*

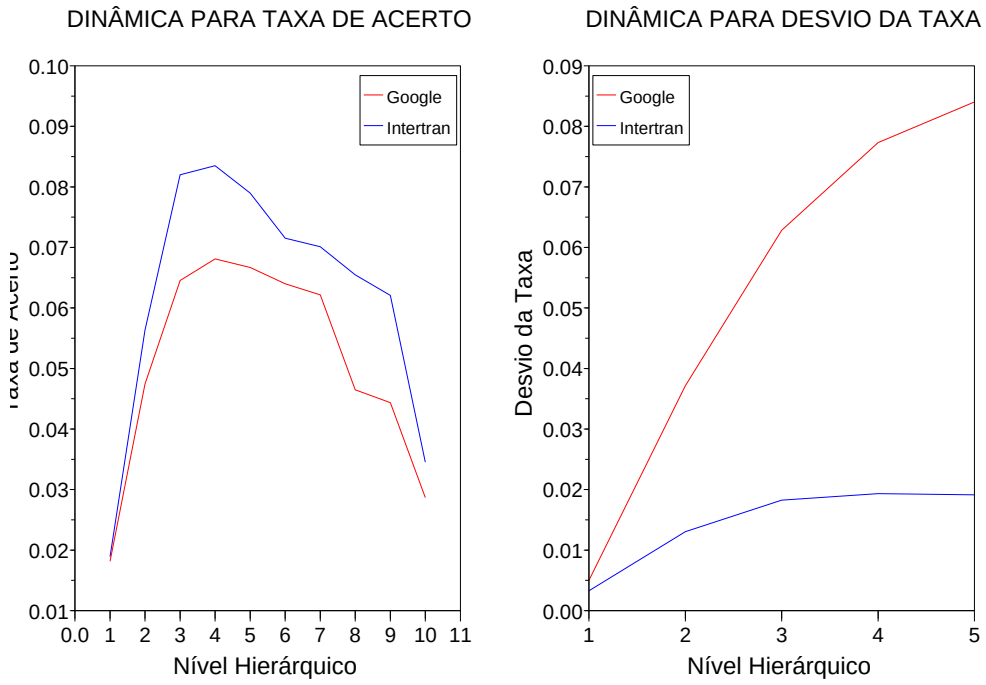

Figura 21: *Variação da taxa de acerto segundo o nível hierárquico utilizado para os tradutores automáticos Google e Intertran, utilizando a definição de erro segundo a equação 39.*

na Tabela 6. Também, utilizou-se o corpus original (sem simplificação) para averiguar as diferenças obtidas de cada simplificação em relação ao corpus não simplificado.

Em um primeira análise de reconhecimento de padrões, buscou-se visualizar como as métricas globais variam entre textos com diferentes graus de simplificação. Assim, foram extraídas para cada texto de cada corpus as principais métricas globais e cada simplificação foi comparada com o texto original, verificando a diferença relativa de uma dada métrica. Seja μ a métrica utilizada, ς a simplificação em questão e ϕ o texto original, então a diferença relativa referente ao par (ς, ϕ) para o texto T, denominado Δ é dado por (40). Os resultados obtidos são ilustrados nas figuras 22, 23 e 24.

$$\Delta_{\varsigma}(t, \mu) = \frac{\mu_{\varsigma}(t) - \mu_{\phi}(t)}{\mu_{\phi}(t)} \tag{40}$$

A Figura 22 exibe os histogramas para cada medida da diferença entre a simplificação menos severa e o texto original. Já neste caso, é possível observar dois padrões que podem distinguir tal simplificação. O primeiro refere-se à medida de grau de entrada (OD) e o segundo ao caminho mínimo 3 (SP_3). No primeiro caso, nota-se uma total distinção, dado que praticamente

Texto Original	Simplif. Natural	Simplif. forte
Quem estava torcendo para o Cristo Redentor se tornar uma das novas maravilhas do mundo tem hoje a última chance de dar uma forcinha para o candidato brasileiro.	Quem estava torcendo para o Cristo Redentor se tornar uma das novas maravilhas do mundo tem hoje a última chance de votar nele.	As pessoas que estão torcendo para o Cristo Redentor se tornar uma das novas maravilhas do mundo tem hoje a última chance de votar nele.
A votação se encerra às 21h pelo horário de Brasília, e o anúncio do resultado ocorre em um megaevento em Lisboa, Portugal, amanhã à noite.	A votação se encerra às 21h de Brasília. O anúncio do resultado ocorre amanhã à noite em um grande evento em Lisboa. Lisboa fica em Portugal.	A votação acaba às 21h de Brasília. O anúncio do resultado ocorre amanhã à noite em um grande evento Lisboa. Lisboa fica em Portugal.

Tabela 6: *Exemplos de simplificações textuais em diferentes níveis. A simplificação natural difere da simplificação do tipo forte pois esta última impõe uma quantidade maior de simplificações.*

todo o histograma está deslocado à direita do zero (diferenciado pela cor azul), indicando que quando um texto é simplificado o seu grau médio tende a aumentar. Especificamente, neste caso, a taxa de aumento está concentrada na faixa de 0% a 15% do texto original. Intuitivamente, tal resultado pode ser explicado pelo fato de que provavelmente um texto simplificado possui a tendência de usar mais vezes uma mesma palavra, além de menos palavras a fim de manter um vocabulário simples e conciso para entendimento. Analogamente, pode-se encontrar também o padrão contrário para as medidas de graus, destacando-se neste caso SP_1 e SP_3, cuja tendência de diminuição global na simplificação pode ser resultado do fato de que em textos simplificados os conceitos (palavras) parecem estar mais próximos para evitar construções linguísticas complexas e desnecessárias.

A Figura 23 apresenta o mesmo conjunto de histogramas que o apresentado na Figura 22, agora para a simplificação tipo forte. Como pode-se notar, ambos os histogramas apresentam o mesmo comportamento, o que está de acordo com a intuição, dado que ambos são simplificações. No entanto, nota-se também que o processo de simplificação mais severo parece estar diretamente correlacionado com tais métricas globais, uma vez que os padrões encontrados anteriormente parecem se confirmar mais intensivamente, principalmente nas métricas de OD,

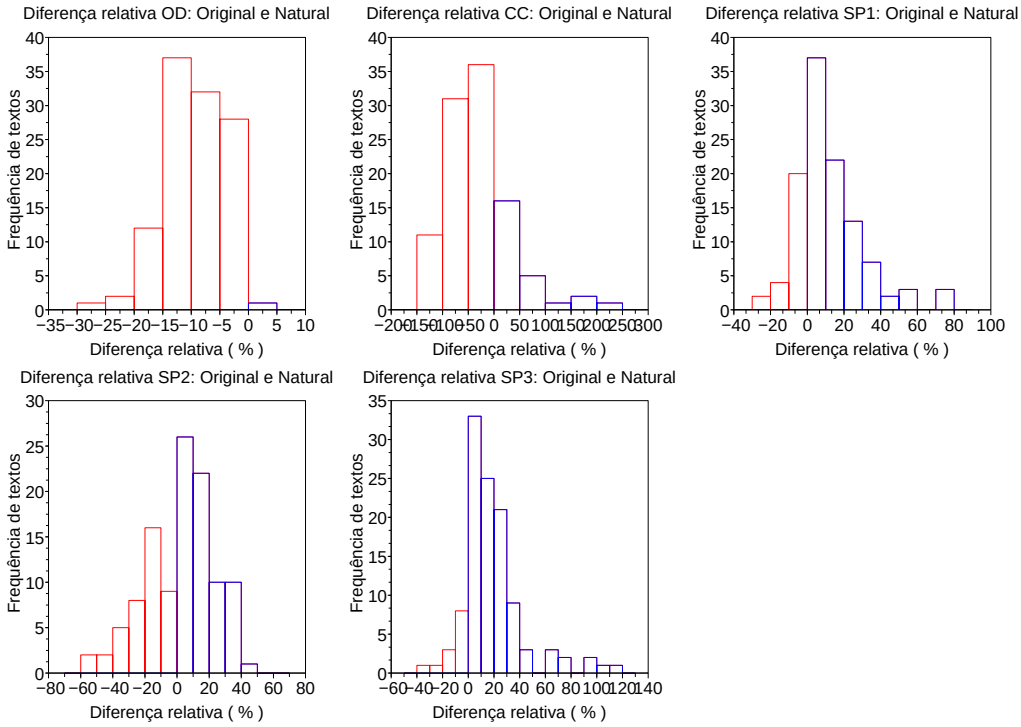

Figura 22: *Histogramas para simplificação natural. As cores das barras estão relacionadas com a posição à esquerda (vermelho) e à direita (azul) do zero.*

SP_1 e SP_3. De fato, os padrões encontrados para simplifações fortes são os mesmos padrões das simplificações naturais, como evidencia a Figura 24 cuja diferença relativa foi feita entre simplificação natural e forte, nesta ordem. Assim, a predominância de barras vermelhas para OD e barras azuis para SP_3 evidenciam que quanto mais severa a simplificação, maior será o seu grau e menor será seu caminho mínimo.

8.3.2 Reconhecimento de padrões segundo o tipo de simplificação para métrica locais

Este experimento apresenta de certa forma as características do experimento descrito na Seção 8.3.1, isto é, buscaram-se padrões nas métricas das redes para distinção entre graus de simplificações. No entanto, aqui foi utilizada uma abordagem local a fim de verificar se um dado vértices da rede original possui correlação com outro vértice da rede simplificada e como esta correlação depende do tipo de simplificação analisada.

O procedimento utilizado foi a comparação das redes plotando as medidas extraídas do texto original e texto simplificado, vértice por vértice. Por exemplo, se na rede original a palavra ρ apresenta para métrica local μ o valor x e se na rede da simplificação a mesma palavra apresenta para a mesma métrica local o valor y então o ponto (x,y) será inserido ao gráfico. Em seguida, dois descritores são extraídos do gráfico: o coeficiente angular e o coeficiente de Pearson (Moore,

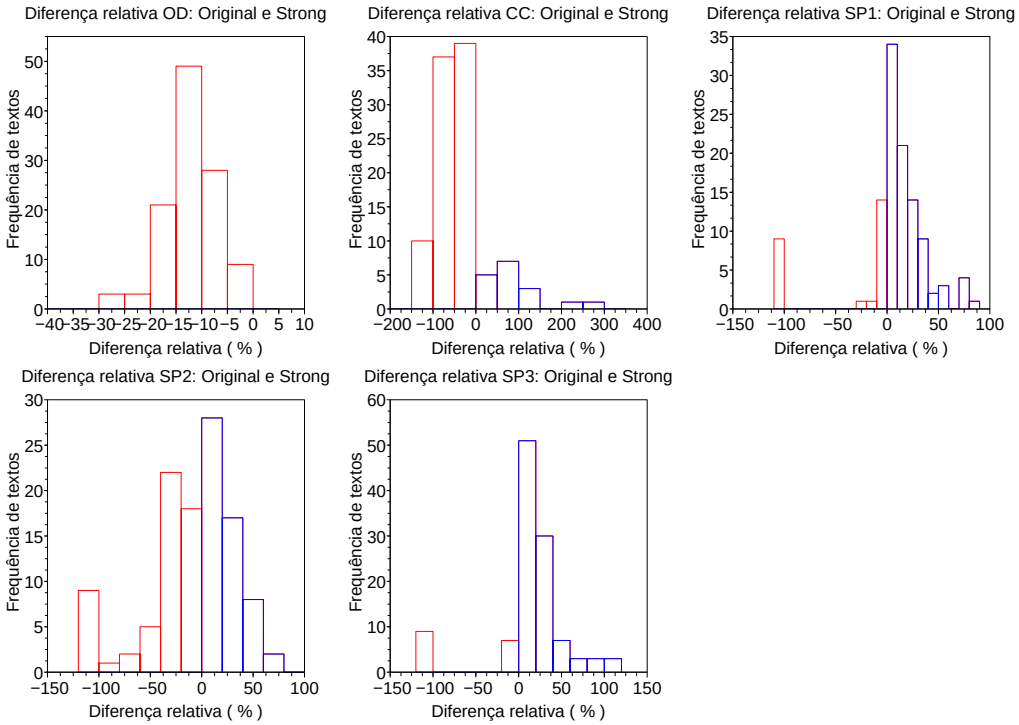

Figura 23: *Histogramas para simplificação forte. As cores das barras estão relacionadas com a posição à esquerda (vermelho) e à direita (azul) do zero.*

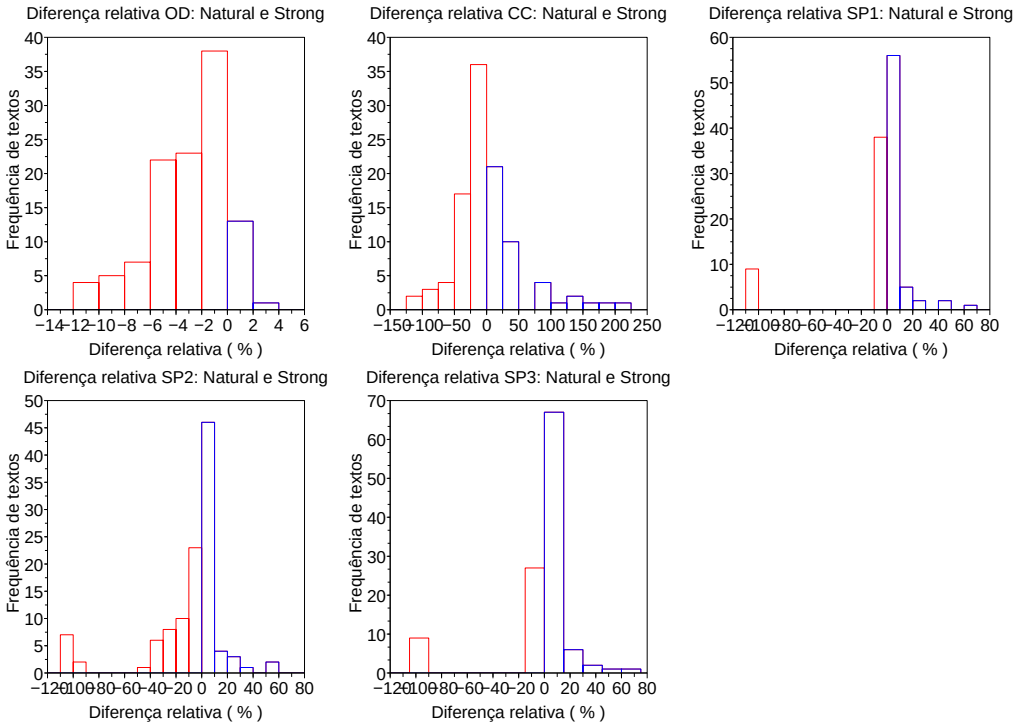

Figura 24: *Histogramas para simplificação forte e natural. As cores das barras estão relacionadas com a posição à esquerda (vermelho) e à direita (azul) do zero.*

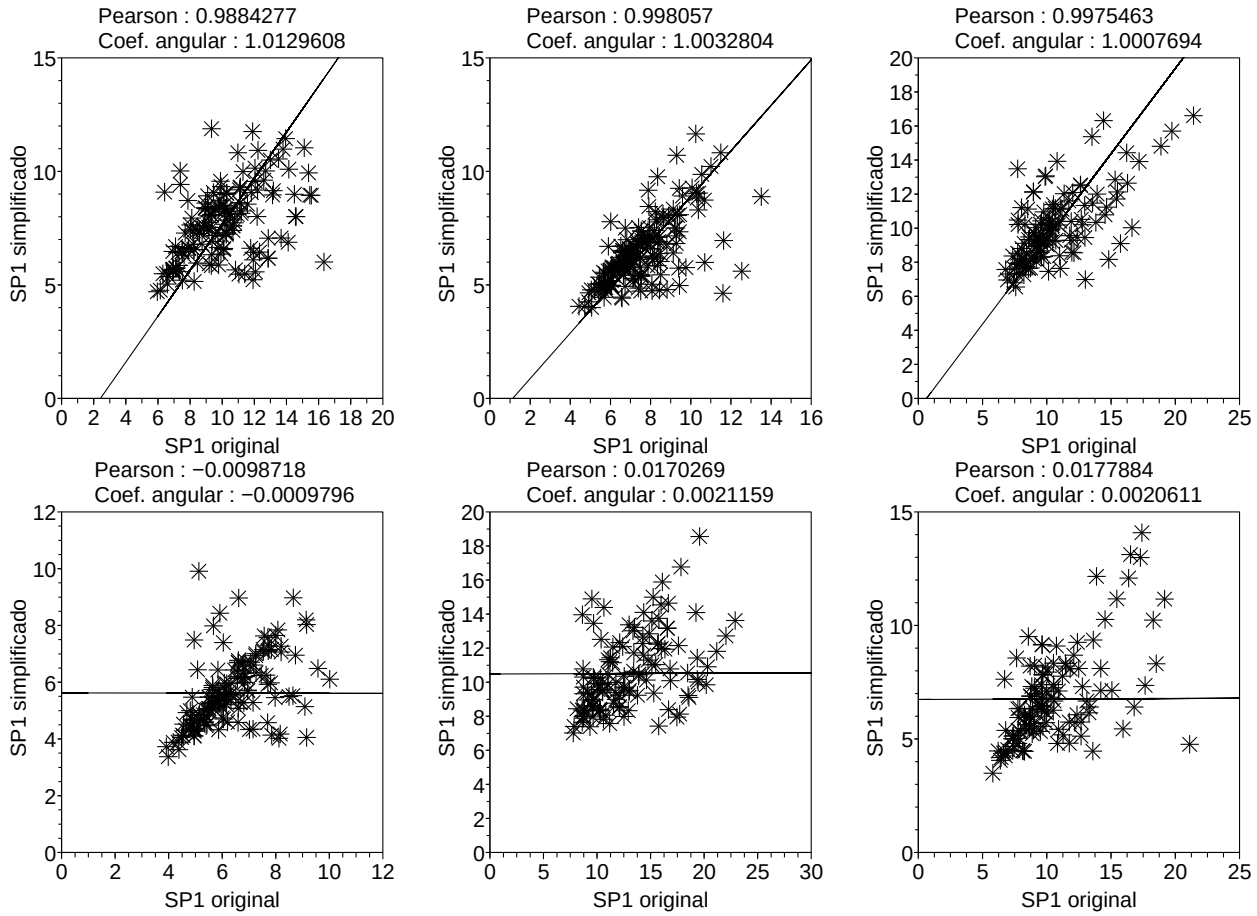

Figura 25: *Exemplo de gráfico para correlação de métricas locais para a métrica de SP_1. Os três gráficos superiores indicam alta correlação e tendência de preservação, com ambos coeficientes próximos de 1. Já os três gráficos inferiores indicam baixa correlação e não preservação, caracterizados pelos baixos valores de coeficientes de Pearson e angular.*

2006), obtidos da melhor reta aproximada pelo método dos mínimos quadrados (Wolberg, 2005). Assim, algumas informações a respeito da preservação de métricas podem ser obtidas tal como se tais redes são similares umas às outras: se existe preservação de métricas locais então os *scatter-plots* deverão apresentar uma característica linear com coeficiente angular próximo de 1, ou seja, com ângulo próximo a 45 graus. Adicionalmente, a quantificação da proximidade em que os pontos se alinham em uma reta é dado pelo coeficiente de Pearson, cujo valor absoluto é próximo de 0 para baixa correlação e próximo de 1 para um alto grau de correlação. A Figura 25 ilustra tais gráficos.

As figuras 26, 27 e 28 apresentam os histogramas obtidos para distribuição de cada coeficiente nas simplificações natural e forte, respectivamente para as métricas CC, OD e SP_1. Para a Figura 26, percebe-se que as distribuições são semelhantes, no entanto com algumas peculiaridades. Quanto à correlação, nota-se que ao se passar da simplificação natural para forte a correlação tende a diminuir, dado que a porção cujo coeficiente de Peasron está en-

tre 0.6 e 1.0 na simplificação natural se distribui para o intervalo compreendido entre 0 e 0.6 na simplificação forte. Ao mesmo tempo, a quantidade de textos cujo coeficiente angular está próximo de 1 diminui na simplificação forte. Este mesmo comportamento parece se repetir para a Figura 28, referente à métrica SP_1, dado que tanto a correlação de Pearson como o coeficiente angular parecem diminuir. Portanto pode-se concluir que o processo de simplificação possui tendência de não preservar métricas de CC e SP_1. Além disso, a linearidade de métricas locais é inversamente proporcional ao grau da simplificação.

A Figura 27, referente aos graus, ilustra que como para a métrica CC, a correlação com uma reta tende a diminuir conforme a simplificação se torna mais intensa, evidenciado pela distribuição de textos no intervalo de 0.9 até 1.0 na simplificação natural para o intervalo de 0.8 até 0.9 na simplificação forte. Apesar desta mudança, a tendência de preservação parece ter mudado pouco, dado que a quantidade de textos com coeficiente angular no intervalo de 0.9 até 1.1 praticamente se mantém. Assim, para os graus pode-se dizer que conforme a simplificação aumenta, sua correlação diminui, apesar de manter a característica de preservação.

Conclui-se que as principais métricas (CC, OD e SP_1) que apresentaram padrões locais podem ser utilizadas para distinguir o grau de simplificação de traduções. Pode-se dizer que em geral, quanto mais simplificado é um texto, menos linear é a relação entre sua respectiva métrica local do texto original (sem simplificação). Também, é possível utilizar o coeficiente angular para distinção, uma vez que as métricas de CC e SP_1 parecem apresentar padrões distintos segundo o tipo de simplificação. Enfim, uma análise multivariada com informação de métricas locais e globais podem evidenciar um excelente conjunto de ferramentas para caracterizar o processo de simplificação textual.

8.3.3 Sumarização e Simplificação por identificação de bordas

Uma importante característica relacionada à estrutura das redes complexas é a identificação de bordas da rede, isto é, dos vértices mais externos. Evidencia-se nesta seção que a definição da diversidade de um nó é capaz de conduzir aos conceitos intuitivos de vértices interiores e vértices exteriores em redes geográficas com estensão direta para a rede de adjacências de palavras. Desta forma, vértices mais exteriores em textos podem ser considerados como de pouca importância, assim abordagens de sumarização padrões (Marcu, 2000) poderiam excluir as sentenças com grande número de vértices exteriores. A definição da diversidade de um nó relaciona-se com a entropia[37] da probabilidade de transição entre vértices vizinhos, através de caminhos aleatórios sobre a rede. Seja i o nó na qual deseja-se calcular a diversidade δ_i, ou seja, deseja-se saber o quão diverso é o acesso deste nó a partir de caminhos aleatórios de tamanho h (o caminho envolve exatamente h vértices), iniciados a partir dos outros (N - 1) vértices da rede. Calcula-se então δ_i como:

[37] Ver Seção 5.1.11.

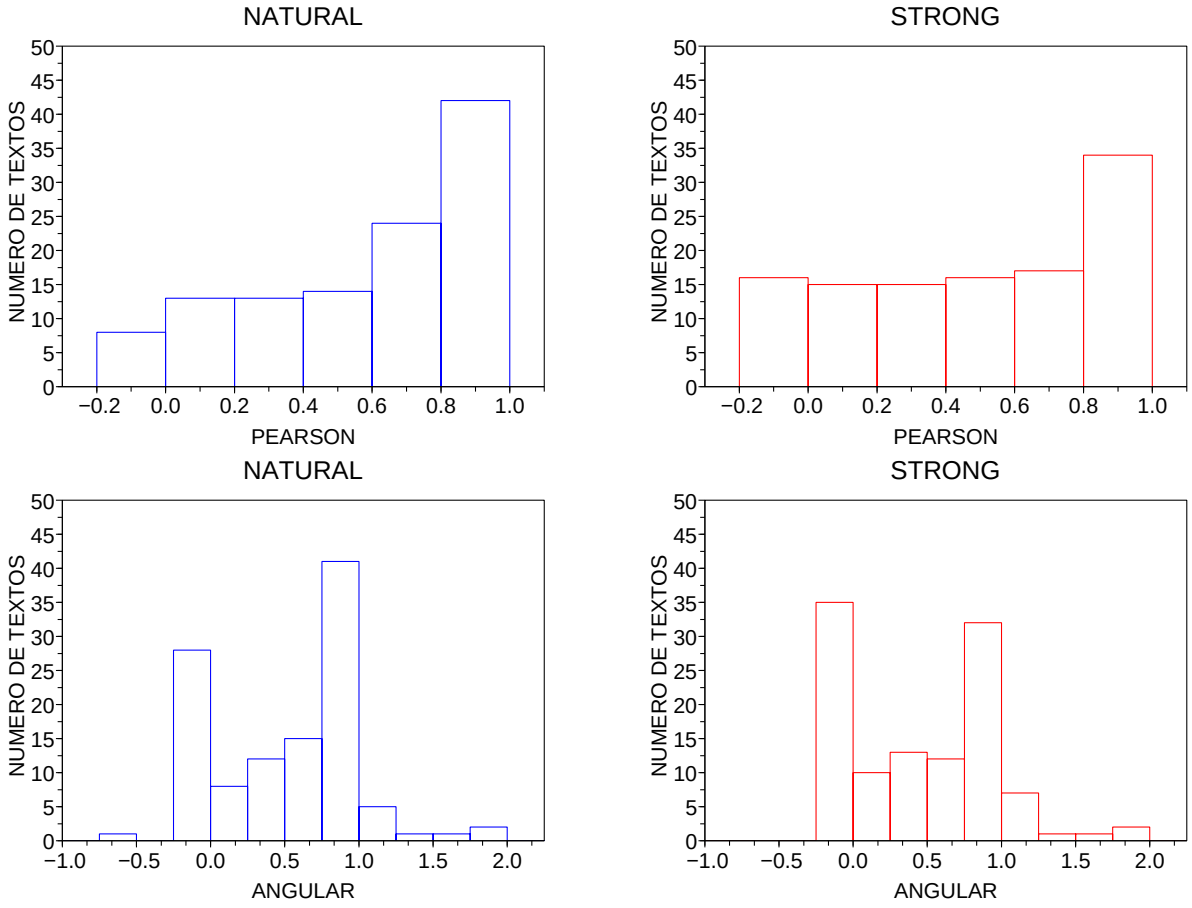

Figura 26: *Comparação de coeficientes obtidos do mapeamento entre texto original e simplificações natural e forte para a métrica CC.*

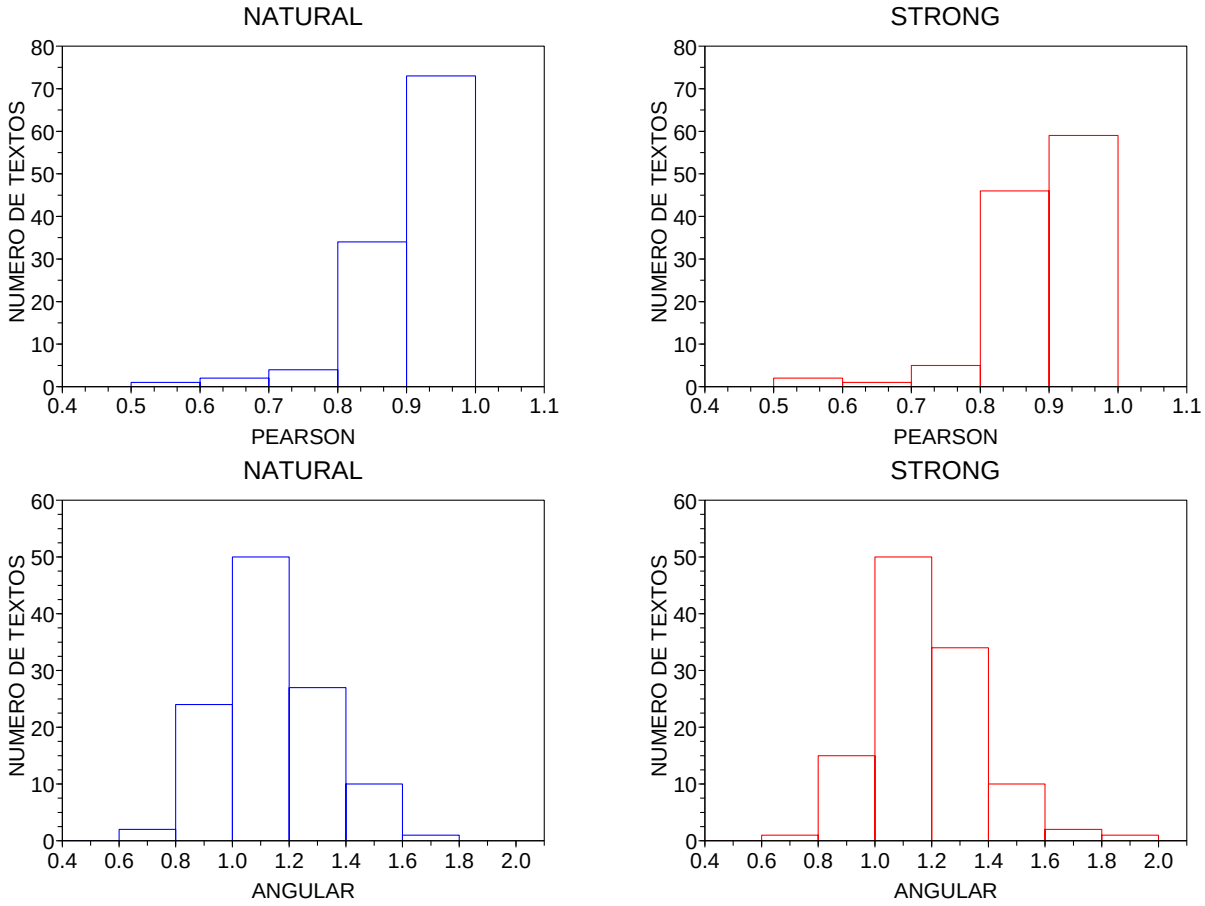

Figura 27: *Comparação de coeficientes obtidos do mapeamento entre texto original e simplificações natural e forte para a métrica OD.*

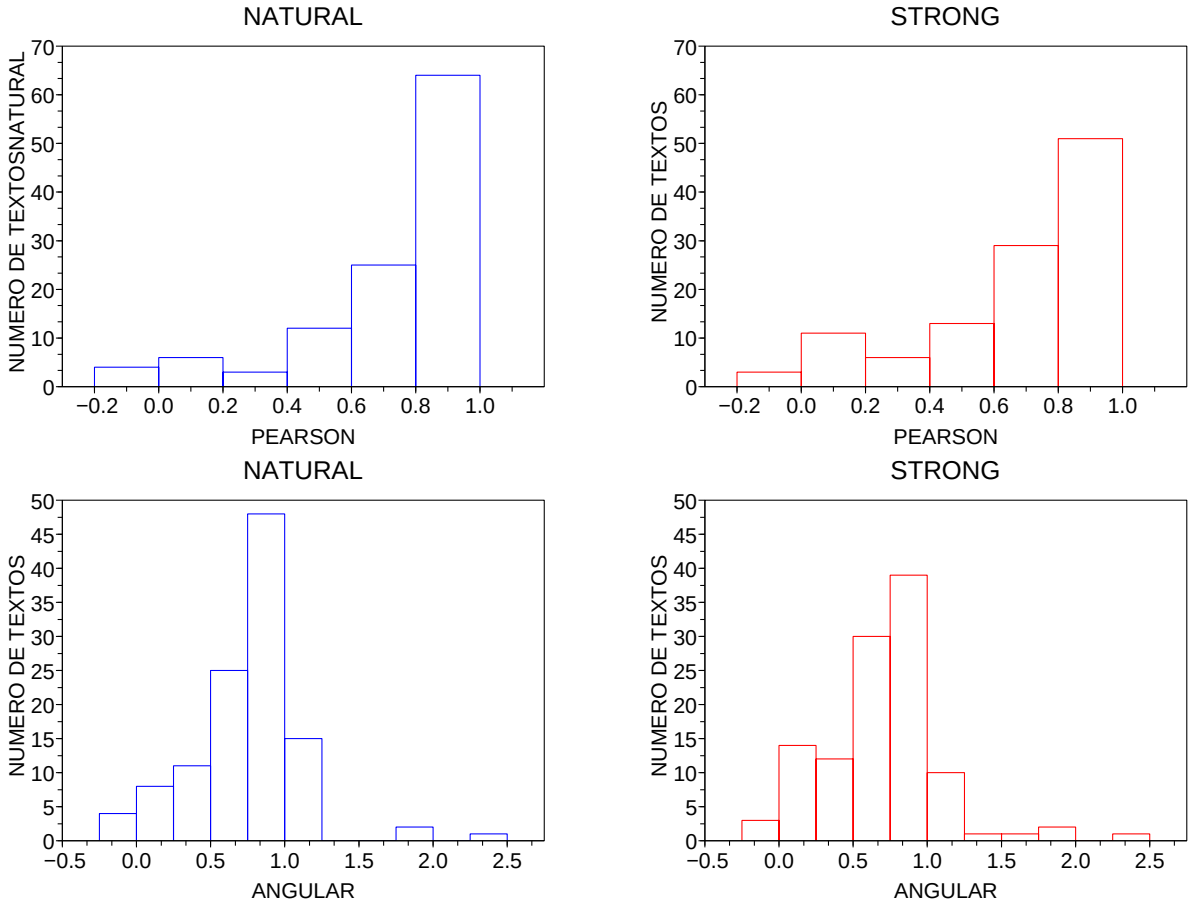

Figura 28: *Comparação de coeficientes obtidos do mapeamento entre texto original e simplificações natural e forte para a métrica* SP_1.

$$\delta_i^h = -\frac{1}{log(N-1)} \sum_{j=1}^{N} P_h(j,i) \cdot log(P_h(j,i)) \tag{41}$$

$$P_h(j,i) = \sum_{c=1}^{NC_h} \prod_{(x,y)\in c} \frac{W_{yx}}{\sum_k W_{kx}} \tag{42}$$

com NC_h representando o número de caminhos de tamanho h desde o vértice i ao vértice j. A equação 42 representa a probabilidade do caminho aleatório seguir a aresta do vértice i ao vértice j. Caso P_h(j,i) seja zero, toma-se o produto no interior da somatória na equação 41 como sendo zero. A fim de verificar a eficiência das equações 41 e 42 na identificação de bordas, estas foram aplicadas em uma rede de 25 vértices, disposta em um quadrado cujo lado envolve 5 vértices, sendo que um vértice na posição não periférica disposto na posição (x,y) possui ligações com seus vizinhos da posição (x+1,y), (x-1,y), (x,y+1) e (x, y-1). A diversidade de cada vértice foi calculada e associada as cores mais claras para os valores maiores e as cores mais escuras para os valores menores. O resultado encontrado é exibido nas figuras 29, 30 e 31, respectivamente para h=1, h=2 e h=3. Pode-se notar que a distribuição das cores está de acordo com a intuição esperada de bordas e esta distribuição se torna mais exata conforme h aumenta. Generalizando tal resultado, a Figura 32 à esquerda ilustra o valor da diversidade para uma rede de 100 vértices análoga ao das figuras 29, 30 e 31, para h=1 até h=6 sendo que o plano xy representa o vértice posicionado na posição (x,y) e o eixo z representa o valor do respectivo nó. Mais uma vez, percebe-se que à medida que se aproxima do centro (5,5), maior o valor da diversidade. A mesma conclusão pode ser deduzida observando o corresponde gráfico de curvas de nível na Figura 32 à direita. Este comportamento acontece porque vértices externos possuem poucas opções de realizar uma caminhada aleatória para acessar outros vértices da rede, enquanto vértices internos possuem mais opções e geralmente estas opções são mais homogêneas, atribuindo a vértices internos maiores valores de diversidade.

A fim de validar esta estratégia com rede de adjacência de palavras, aplicou-se as mesmas equações a textos pertencentes a um corpus representando por 300 redações do ENEM relacionadas ao tema da importância do voto. A estratégia verificada aqui consistiu em averiguar para cada texto e para cada valor de h (de 1 até 3), quais eram as palavras (vértices da rede) que apresentavam os 10% maiores valores das diversidade para cada nível. Desta forma, para cada texto, uma dada palavra pode aparecer até h vezes entre os 10% maiores valores, já que este cálculo é realizado para cada nível. Estes valores são sumarizados na Tabela 7. Nota-se que várias palavras de grande importância relacionada ao tema aparecem na tabela (destacado), além das palavras de uso comum na linguagem, como os verbos **ser**, **ter** e **fazer**. A fim de confrontar este resultado, a Tabela 8 exibe as palavras de maior frequência entre as 10 % de menor diversidade para cada texto. De fato, palavras com pouca relação com o tema apare-

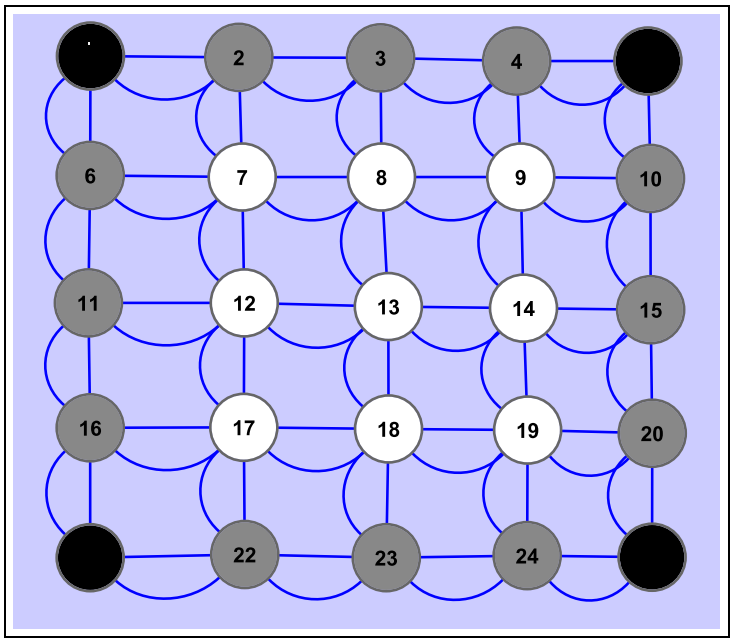

Figura 29: *Identificação de bordas utilizando a diversidade de acesso ao vértice para h = 1. Quanto mais escuro, menor é o valor da diversidade e mais próximo da borda o vértice se apresenta.*

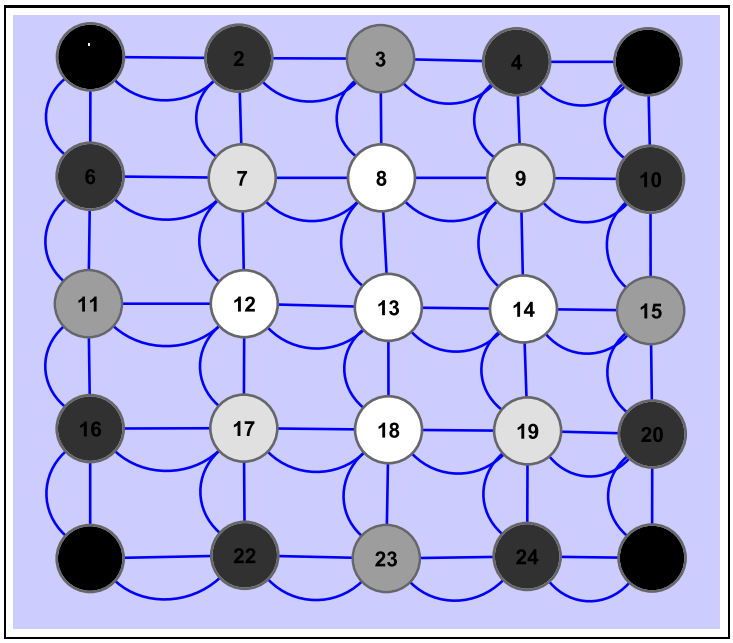

Figura 30: *Identificação de bordas utilizando a diversidade de acesso ao vértice para h = 2. Quanto mais escuro, menor é o valor da diversidade e mais próximo da borda o vértice se apresenta.*

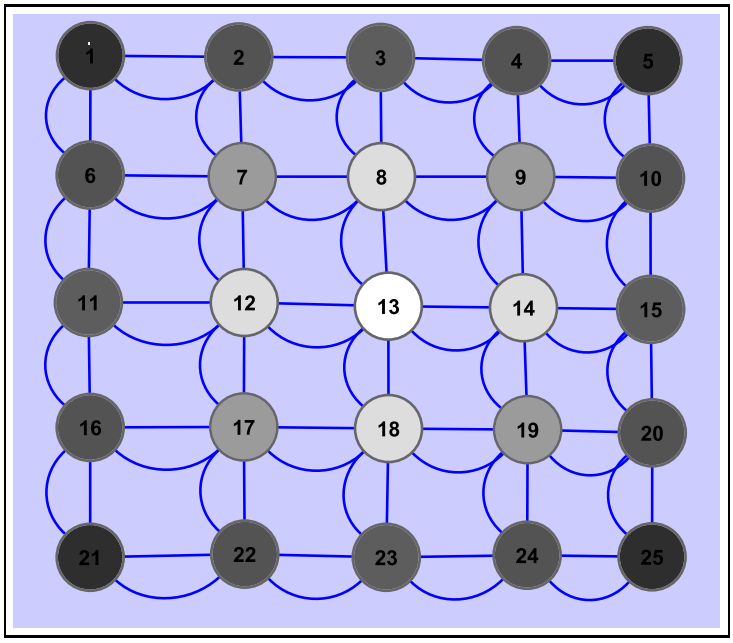

Figura 31: *Identificação de bordas utilizando a diversidade de acesso ao vértice para h = 3. Quanto mais escuro, menor é o valor da diversidade e mais próximo da borda o vértice se apresenta.*

cem frequentemente como bordas da rede[38]. Neste ponto, vale a pena ressaltar que o nível de interioridade tem pouca relação com as medidas padrões, como os graus, como pode ser visto na Figura 33: embora para h=1 as métricas de OD e ID apresentarem alta correlação com a diversidade, torna-se claro que à medida que h aumenta (e consequentemente a distinção das bordas se torna mais aprimorada) a correlação cai, logo esta forma de caracterização não é redundante na especificação de palavras chaves.

Pode-se perceber que este método tem grande tendência em eleger as palavras chaves do texto, favorecendo a estratégia de sumarização com fins de simplificação usando a técnica denominada de extração de palavras chaves. Tal método é um dos mais difundidos no que se refere à tarefa de sumarização (Souza e Nunes, 2001) e sua aplicação é a seguinte: dado um conjunto de palavras-chaves relativas a um texto, todas as sentenças que apresentarem pelo menos uma palavra chave são utilizados no texto simplificado por resumo, sendo as outras descartadas. Tal método insere-se na sumarização extrativa, dado que o resumo possui sentenças pré-existentes no texto original. Adicionalmente, outras abordagens para sumarização com fim de simplificação podem ser utilizadas, como por exemplo a seleção da quantidade de palavras chaves de acordo com a porcentagem de resumo esperada. Também é possível utilizar a modelagem descrita em Antiqueira et al. (2009), na qual vértices correspondem às sentenças e as arestas são criadas entre duas sentenças se estas possuem duas palavras em comum. A partir desta

[38]Quatro palavras da Tabela 7 apareceram na Tabela 8 (ocultas nesta última) mostrando que em algumas redações (talvez aquelas com notas inferiores) tais palavras podem aparecer como bordas.

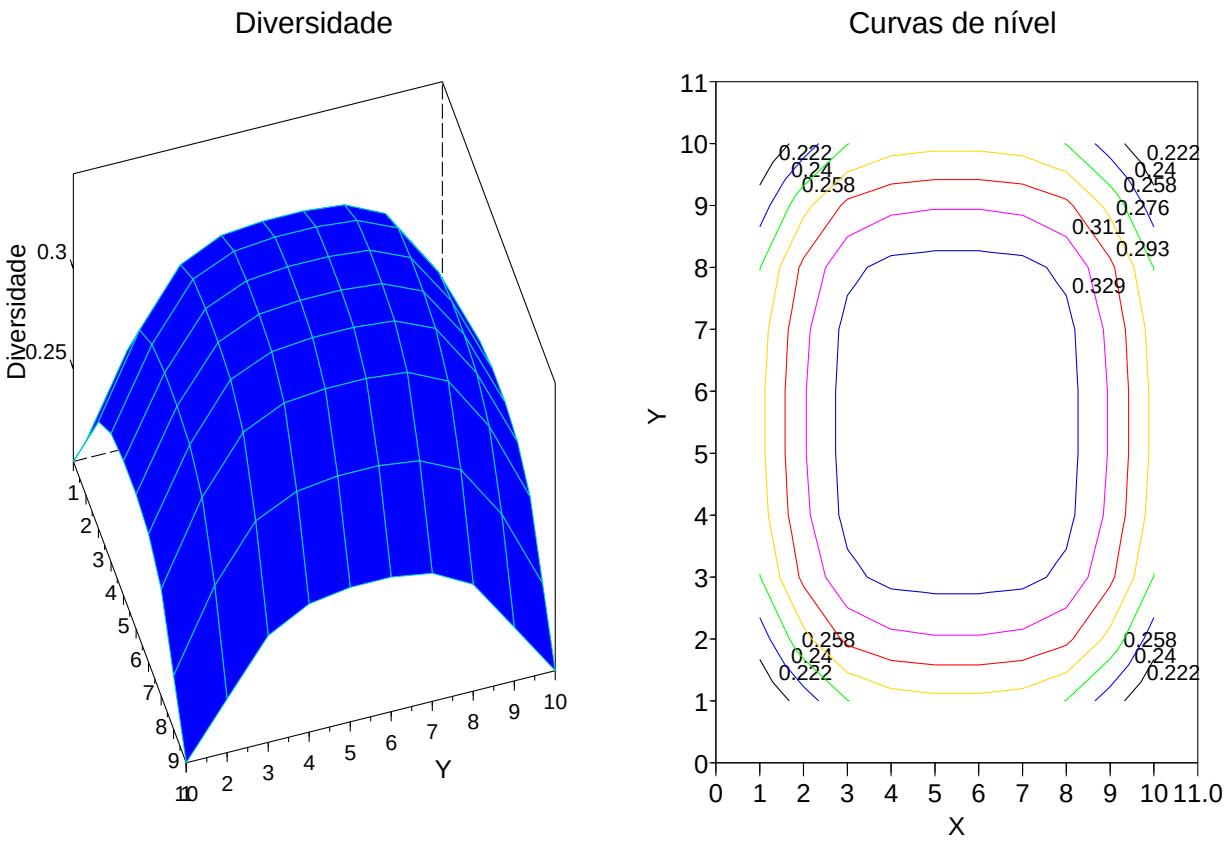

Figura 32: *Diversidade em função da posição do vértice em forma de função de duas variáveis (esquerda) com suas respectivas curvas de nível (direita) evidenciam que vértices internos possuem maiores valores da métrica de diversidade da entropia.*

Palavra	Frequência	Palavra	Frequência
ser	732	fazer	132
não	447	pessoa	104
votar	407	**candidato**	104
ter	384	estar	90
voto	379	ir	80
poder	255	**político**	84
país	246	melhor	63
direito	246	**brasileiro**	58
dever	162	**povo**	58
Brasil	162	saber	57

Tabela 7: *Tabela representando o número de vezes que uma dada palavra foi considerada entre as 10% mais importantes pela estratégia de detecção de bordas.*

Palavra	Frequência	Palavra	Frequência
ano	125	agora	62
bem	99	antes	61
cidadão	82	analisar	61
ajudar	74	conquista	60
acabar	72	arma	52
bom	72	acontecer	52
ainda	66	através	50
acreditar	64	coisa	45

Tabela 8: *Tabela representando o número de vezes que uma dada palavra foi considerada entre as 10% com menor valor de diversidade, representando portanto as bordas.*

modelagem, segue diretamente o conceito de sentença-chave em analogia às palavras-chaves referidas.

9 Conclusão

Métricas padrões e hierárquicas retiradas das redes complexas resultantes da modelagem de textos (traduções automáticas e simplificações) foram empregadas para reconhecimento de padrões de qualidade de traduções e de níveis de complexidade textual, além de serem usadas em um algoritmo de reconhecimento de palavras-chaves para sumarização. Com respeito às traduções automáticas, percebeu-se que quanto maior a qualidade, maior é a tendência de preservação (essencialmente para o par espanhol-português) de uma série de métricas combinadas em níveis hierárquicos. Para o tema de simplificação textual os experimentos mostraram forte correlação entre as métricas de OD e SP_1 e SP_3, através da alta tendência de aumentar para OD e diminuir para SP_1 e SP_3 conforme o texto se torna mais simplificado. Adicionalmente, foi mostrado que a diversidade de caminhos através de um vértice da rede é capaz de sintetizar o conceito

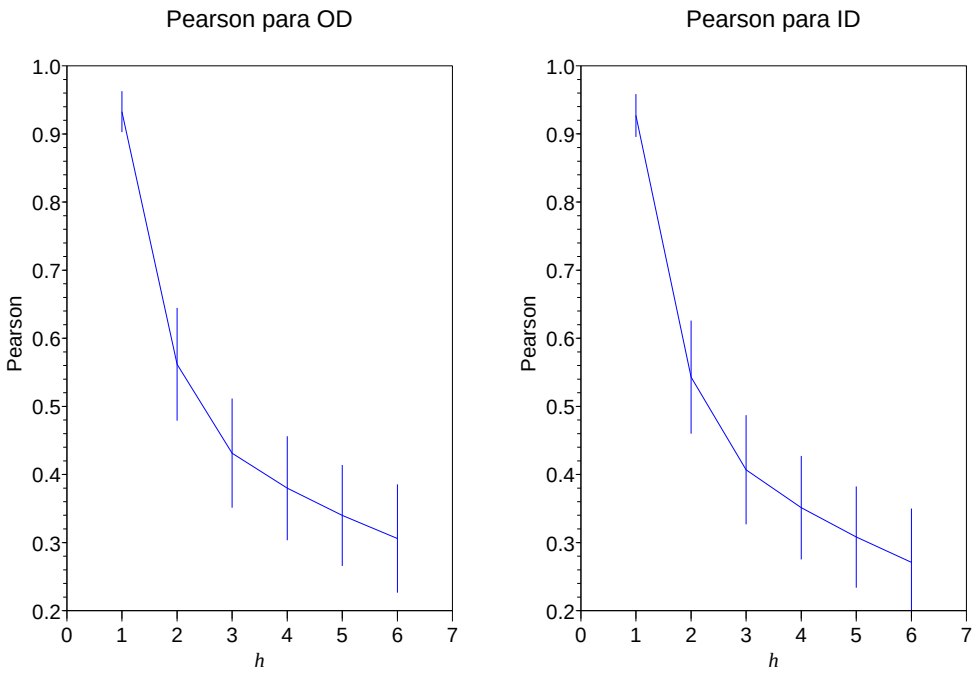

Figura 33: *Dinâmica da correlação da diversidade com os graus de entrada e saída, de acordo com o nível h.*

de palavra-chave, útil na tarefa de sumarização visando a simplificação.

Para os trabalhos futuros, pretende-se verificar como se comporta a tradução humana quanto à dinâmica da taxa de acerto sobre os níveis hierárquicos e se este comportamento é condizente com alguns dos tradutores automáticos analizados. Já para o tema de simplificação textual pretende-se averiguar os padrões encontrados e validá-los através da correlação com medidas já existentes de níveis de complexidades a fim de gerar através do uso de aprendizado de máquina uma métrica própria baseada exclusivamente ou parcialmente em redes complexas. Também pretende-se aprofundar no assunto referente à extração de palavras chaves para o desenvolvimento de sumarizadores automáticos a fim de compará-los com os existentes na literatura.

10 Referências

AIRES, R.V.X.; ALUÍSIO, S.M.; KUHN, D.C.S.; ANDREETA, M.L.B.; OLIVEIRA JR., O.N. Combining Multiple Classifiers to Improve Part of Speech Tagging: A Case Study for Brazilian Portuguese. In the Proceedings of the Brazilian AI Symposium (SBIA'2000), 20-22, 2000.

ALBERT, R.; BARABÁSI, A.L. Statistical Mechanics of Complex Networks, Rev. Modern Phys., 74, 47-97, cond-mat/0106096, 2002.

AMANCIO, D.R.; ANTIQUEIRA, L.; PARDO, T.A.S.; COSTA, L.F.; OLIVEIRA JR., O.N.; NUNES, M.G.V. Complex networks analysis of manual and machine translations. International Journal of Modern Physics C, 19(4):583-598, 2008.

ANTIQUEIRA, L.; OLIVEIRA JR., O. N.; COSTA, L.F.; NUNES, M. G. V. A Complex Network Approach to Text Summarization , Information Sciences, 179(5), 584-599, 2009.

ANTIQUEIRA, L; NUNES, M. G. V.; OLIVEIRA Jr., O. N.; COSTA, L.F. Strong correlations between text quality and complex networks features. Physica A, 373:811-820, 2007.

ANTIQUEIRA, L.; PARDO, M. G. V. ; NUNES, M. G. V.; OLIVEIRA JR., O. N.; COSTA, L.F. Some issues on complex networks for author characterization. In Proceeedings of the Workshop in Information and Human Language Technology (TIL'06), 2006.

BARTHÉLEMY, M; BARRAT, A.; PASTOR-SATORRAS, R.; VESPIGNANI, A. Characterization and modeling of weighted networks. Physica A, 346:34-43, 2005.

BATES, M.Models of natural language understanding. Proceedings of the National Academy of Sciences of the United States of America, Vol. 92, No. 22 (Oct. 24, 1995), pp. 9977-9982, 1995.

BICK, E. The Parsing System PALAVRAS: Automatic Gramatical Analysis of Porutugese in a Constraint Grammar Framework. Aarhus University Press, 2000.

BOITET, C. (Human-Aided) Machine Translation: A Better Future ? Grenoble, 1994.

CALDEIRA, S.M.G.; PETIT LOBÃO, T. C.; ANDRADE, R. F. S.; NEME, A.; MIRANDA; J. G. V. The network of concepts in written texts. European Physical Journal B, 49:523-529, 2006.

CANCHO, R.F.; SOLÉ, R.V. The Small World of Human Language. Proceedings of The Royal Society of London. Series B, Biological Sciences, 268, 2261-2265, 2001.

CARLETTA, J. Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics, 22(2), pp. 249-254, 1996.

CASELI, H.M; NUNES, M.G.V. Alinhamento Sentencial e Lexical de Corpus Paralelos: Recursos para a Tradução Automática. Estudos Lingüísticos, 34, 356-361, 2005.

COHEN, W. W. Fast Effective Rule Induction. Proc. of the 12th International Conference on Machine Learning. Tahoe City, Califórnia, 1995.

CORMEN, T.H.; LEISERSON, C.E.; RIVEST, R.L.; STEIN, C. Introduction to Algorithms, The MIT Press, 2002.

COSTA, L.F.; OLIVEIRA JR., O.N.; TRAVIESO, G.; RODRIGUES, F.A.; VILLAS BOAS, P.R.; ANTIQUEIRA, L; VIANA, M.P.; ROCHA, L.E.C. Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. Physics and Society, 2009.

COSTA, L.F.; ANDRADE, R.F.S. What are the Best Hierarchical Descriptors for Complex Networks ? New J. Phys. 9 311, 2007.

COSTA, L.F.; ROCHA, L.E.C. A generalized approach to complex networks European. Physical Journal B, v. 50, p. 237-242, 2006.

COSTA, L.F. The Hierarchical Backbone of Complex Networks. Phys. Rev. Lett. 93: 098702, 2004.

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

DOROGOVTSEV, S. N.; MENDES, J. F. F. Language as an evolving word web. Proceedings of the Royal Society of London B, 268:2603, 2001.

EDMONDS, P. Choosing the word most typical in context using a lexical co-occurrence network. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics,

pages 507-509, 1997.

FLESCH, R. A new readability yardstick. Journal of Applied Psychology, Vol. 32, pp. 221-233, 1948.

FURNKRANZ, J.; WIDMER, G.Incremental Reduced Error Pruning. Internacional Conference on Machine Learning, 1994.

COSTA, L. F.; OLIVEIRA JR., O. N.; TRAVIESO, G.; Rodrigues, F. A.; VILLAS BOAS, P. R.; ANTIQUEIRA, L.; VIANA, M. P.; ROCHA, L. E. C. Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. Physics and Society, 2008.

HUTCHINS, W.J.; SOMERS, H.L. An Introduction to Machine Translation. London: Academic Press, 1992.

KINCAID, J. P.; FISHBURNE, R. P., Jr.; ROGERS, R. L.; CHISSOM, B. S. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel, Research Branch Report 8-75, Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN, 1975.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12): 1137-1143, 1995.

MARCU, D. The Theory and Practice of Discourse Parsing and Summarization, The MIT Press, A Bradford Book, 2000.

MARGARIDO, P.R.A.; NUNES, M.G.V.; PARDO, T.A.S.; OLIVEIRA JR., O.N. Redes complexas para processamento de língua natural: um estudo sobre a estabilização de métricas das redes. Revista Eletrônica de Iniciação Científica - REIC, Ano VIII, N. 3, 2008.

MARTINS, R.T.; HASEGAWA, R.; NUNES, M.G.V.; MONTILHA, G.; OLIVEIRA JR., O.N. Linguistic issues in the development of ReGra: a Grammar Checker for Brazilian Portuguese.Natural Language Engineering, Volume 4, p287-307;Cambridge University Press, 1998.

MAX, A. Writing for Language-impaired Readers. In the Proceedings of Seventh International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), pp. 567-

570, Mexico City, Mexico, 2006.

MAZIERO, E.G.; PARDO, T.A.S; ALUÍSIO, S.M. Ferramenta de Análise Automática de Inteligibilidade de Córpus (AIC). Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional. NILCTR-08-08, 2008.

MIHALCEA, R. Random walks on text structures. In Proceedings of CICLing-2006, volume 3878 of LNCS, pages 249-262, February, 2006.

MILGRAN, S. The small world problem. Psychology Today, 1(1):60-67, 1967.

MILLER, G.A.Wordnet: a dictionary browser. Proceedings of the First International Conference on Information in Data. University of Waterloo, 1985.

MOORE, D. Basic Practice of Statistics. WH Freeman Company, pp 90-114, ISBN 0716774631, 2006.

MOTTER, A.E.; MOURA, A.P.S.; LAI, Y.C.; DASGUPTA, P. Topology of the Conceptual Network of Language, Phys. Rev. E, 65, 065102, 2002.

NEWMAN, M.E.J.The Structure and Function of Complex Networks. SIAM Review 45, 167-256, cond-mat/0303516, 2003.

NUNES, M.G.V. et al. O Processo de Construção de um Léxico para o Português do Brasil: Lições Aprendidas e Perspectivas. II Encontro para o Processamento Computacional de Português Escrito e Falado, p.61-70. CEFET-PR, Curitiba, 1996.

OLIVEIRA JR., O. N.; MARCHI, A. R.; MARTINS, M. S.; MARTINS, R. T. A Critical Analysis of the Performance of English-Portuguese-English MT Systems. V Encontro para o processamento computacional da Lingua Portuguesa Escrita e Falada (PROPOR 2000) Atibaia, SP, 20 a 22 Novembro 2000.

PANG, B.; LEE, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the ACL 2004, pages 271-278, 2004.

PAPINEI, K.; ROUKOS, S.; WARD, T.; ZHU, W. J. BLEU: a method for automatic evaluation of machine translation"in ACL-2002: 40th Annual meeting of the Association for Computati-

onal Linguistics pp. 311-318, 2002.

PARDO, T.A.S; ANTIQUEIRA, L.; NUNES, M. G. V.; OLIVEIRA JR., O. N.; COSTA, L.F. Using complex networks for language processing: The case of summary evaluation. In Proceedings of the International Conference on Communications, Circuits and Systems (ICCCAS'06) - Special Session on Complex Networks, 2006.

QUINLAN, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

RATNAPARKI, A. A Maximum Entropy Part-Of-Speech Tagger. In the Proceedings of the Empirical Methods in Natural Language Processing Conference, University of Pennsylvania, 1996.

SIGMAN, M.; CECCHI, G.A. Global Organization of the Wordnet Lexicon. Proceedings of the National Academy of Sciences, 99, 1742-1747, 2002.

SOUZA, C.F.R., NUNES, M.G.V. Avaliação de Algoritmos de Sumarização Extrativa de Textos em Português. Relatório Técnico do ICMC-USP. NILCTR-01-09, 2001.

SPRINZAK, E.; MARGALIT, H. Correlated sequence-signatures as markers of protein-protein interaction. J. Mol. Biol. 311, 681, 2001.

VERONIS, J. Parallel Text Processing: Alignment and Use of Translation Corpora, capítulo 1, 1-24, 2000.

WATTS, D.J. Small worlds: the dynamics of networks between order and randomness. Princeton University Press, 1999.

WITTEN, I. H.; FRANK, E. Data Mining: Pratical machine learning tools and techniques. Morgan Kauffmann, 2005.

WOLBERG, J. Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments, Springer, 2005.