Redes complexas para processamento de língua natural: um estudo sobre a estabilização de métricas das redes

Paulo R. A. Margarido[1], Maria das Graças V. Nunes[1], Thiago A. S. Pardo[1], Osvaldo N. de Oliveira Jr.[2]

[1]Instituto de Ciências Matemática e de Computação (ICMC)
Universidade de São Paulo (USP) – São Carlos, SP – Brasil

[2]Instituto de Física de São Carlos (IFSC)
Universidade de São Paulo (USP) – São Carlos, SP – Brasil

pauloram@grad.icmc.usp.br, {gracan,taspardo}@icmc.usp.br, chu@ifsc.usp.br

Abstract. *An extension to graph theory, the theory of complex networks has been recently developed, and several areas of knowledge have begun to employ complex networks to a wide variety of tasks with positive results, including the area of Natural Language Processing. However, this association between complex networks and natural language processing aplications has not been thoroughly explored, since the limits and capabilities of complex networks as a means for representing texts are somewhat unknown. The focus of this work is to determine how stable the complex networks are when the metrics used to evaluate them are considered.*

Resumo. *Uma extensão da teoria dos grafos, a teoria das Redes Complexas foi recentemente formulada, e diversas áreas começaram a empregar as redes complexas para várias tarefas com resultados positivos, inclusive a área de Processamento de Língua Natural. Porém, esta associação entre redes complexas e aplicações de processamento de língua natural ainda não foi devidamente explorada, uma vez que não se conhecem bem os limites e capacidades das redes enquanto representação de textos. O foco deste trabalho é determinar quão estáveis as redes complexas são quando as métricas usadas para avalia-las são consideradas.*

1. Introdução

As redes complexas têm recentemente recebido grande visibilidade em várias áreas da computação devido à sua flexibilidade para representar e lidar com diversos tipos de problemas. Com o desenvolvimento de uma representação de textos por meio de redes complexas, permitiu-se que várias aplicações clássicas de PLN (Processamento de Língua Natural) fossem realizadas utilizando as redes complexas [4], por exemplo, as áreas de atribuição de autoria [3], avaliação de qualidade [2], sumarização automática e tradução automática, entre outras, com resultados comparáveis ao estado da arte [7].

Já vem se fazendo pesquisas sobre a aplicabilidade das redes nestas tarefas, porém, como é uma pesquisa iniciada há relativamente pouco tempo, as informações obtidas ainda são limitadas. O foco deste trabalho é tentar obter mais informações sobre a dinâmica das redes complexas enquanto representação de textos para aplicações de

PLN [1], a fim de desenvolver uma base sólida sobre a qual futuros experimentos possam ser realizados com uma maior precisão, determinando quais métricas das redes complexas são mais previsíveis e fornecem resultados mais garantidos. Outro objetivo deste trabalho é determinar um limite inferior confiável para o tamanho dos textos modelados como redes para que as métricas sejam previsíveis o bastante. Este tipo de informação é importante para determinar quais aplicações são as mais recomendáveis para as redes complexas, contribuindo assim para que se ache um nicho de aplicação com redes complexas em PLN. Com base nessa discussão, realiza-se um experimento com a tarefa modelagem de autoria.

Na Seção seguinte está uma breve definição das principais características das redes complexas e de como textos são representados neste trabalho. Na Seção 3 estão descritas as métricas utilizadas no experimento que conduzimos, enquanto, na Seção 4, encontram-se informações sobre o experimento, como o corpus com que trabalhamos, a metodologia empregada e os resultados observados. Finalmente, as conclusões se encontram na Seção 5.

2. Redes complexas e representação textual

As redes normalmente são usadas quando os conceitos da teoria dos grafos não permitem obter as informações desejadas, uma vez que as redes complexas têm algumas características que as favorecem neste tipo de aplicação. O traço mais importante das redes complexas é que se trata de grafos com alto número de vértices ou nós, o que também ocorre em problemas reais. Algumas outras características também são relevantes:

- Apesar do tamanho das redes, o número de nós entre um dado par de nós é relativamente pequeno, isto é, o caminho mínimo entre dois nós tem baixo custo; tais redes são conhecidas como redes mundo pequeno ou *small world* [7].
- Quando um grupo pequeno em relação ao tamanho da rede possui nós que são fortemente conectados entre si, trata-se de um aglomerado (ou *cluster*). As redes complexas têm uma tendência a ter um alto coeficiente de aglomeração (definido na Seção 3), isto é, elas tendem a apresentar pequenos grupos fortemente conectados.
- Outra característica marcante das redes complexas é o fato de que elas são livres de escala ou *scale-free*. Isso significa que redes complexas têm poucos *hubs* (nós com grande quantidade de conexões) e muitos nós com baixa quantidade de conexões, e novos nós adicionados tendem a se ligar a nós já bem conectados.

Várias destas características podem ser observadas nas Figuras 1-2. A primeira é uma rede criada a partir de um texto pequeno: as características da rede não são tão visíveis quanto na segunda figura, que é um mapeamento de uma rede social como rede complexa. Na segunda, existem alguns vértices extremamente conectados, ou seja, *hubs*, assim como é possível observar a existência de diversos *clusters* na rede.

Neste trabalho, antes de se modelar um texto como uma rede complexa [4], uma etapa de pré-processamento é necessária. Nesta etapa, primeiramente são removidas as *stopwords* (palavras muito freqüentes e com pouco significado) e, em seguida, as palavras remanescentes são lematizadas, isto é, flexionadas em uma forma comum (por exemplo, todos os verbos são flexionados para o infinitivo) para evitar problemas com dados esparsos. Cada uma dessas palavras será um nó na rede, como na Figura 3.

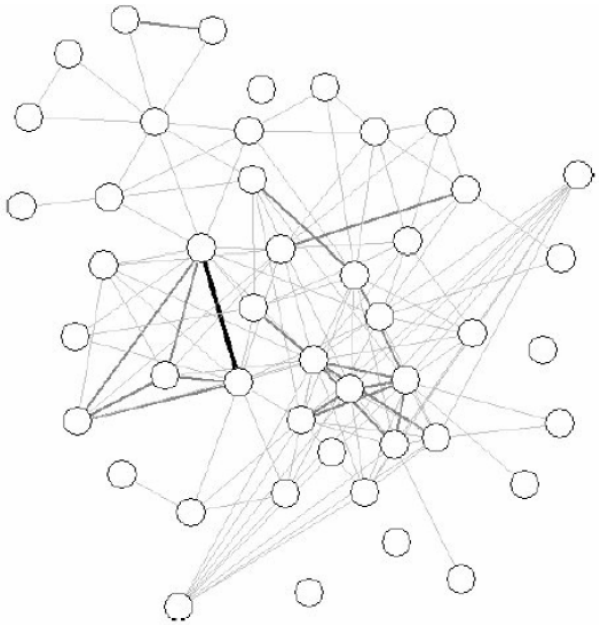


Figura 1. Exemplo de um pequeno texto modelado como rede complexa.

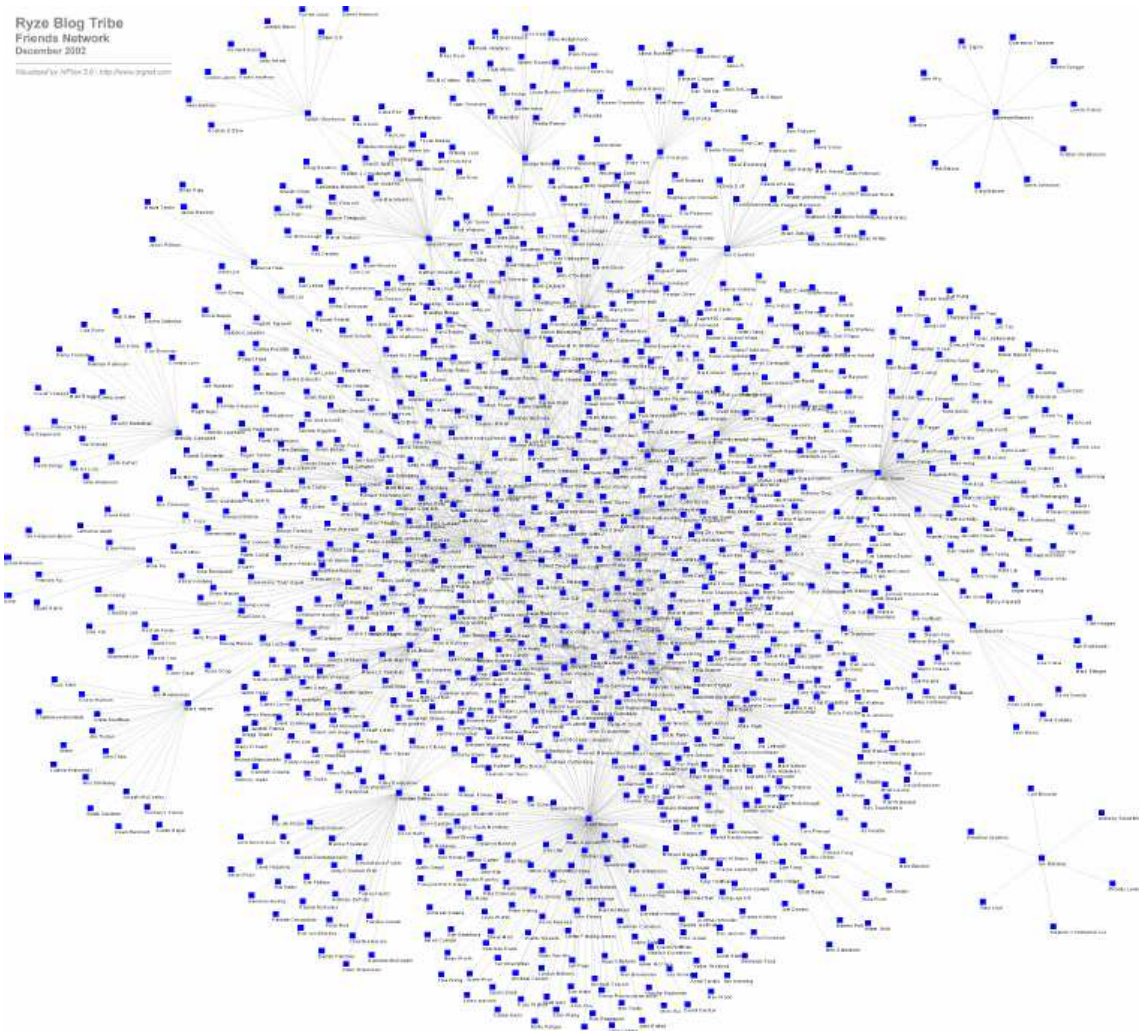
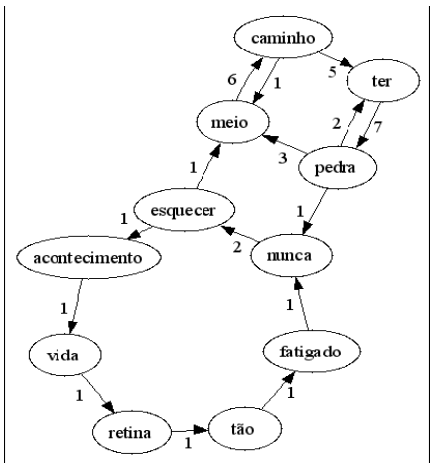


Figura 2. Exemplo de rede social modelada como rede complexa.



No meio do caminho tinha uma pedra
tinha uma pedra no meio do caminho
tinha uma pedra
no meio do caminho tinha uma pedra.

Nunca me esquecerei desse acontecimento
na vida de minhas retinas tão fatigadas.
Nunca me esquecerei que no meio do caminho
tinha uma pedra
tinha uma pedra no meio do caminho
no meio do caminho tinha uma pedra.

Carlos Drummond de Andrade

Figura 3. Exemplo de rede complexa e do texto correspondente pela modelagem aqui proposta

As redes dos textos utilizados neste artigo são então criadas da seguinte maneira: cada associação entre palavras no texto gera uma aresta direcionada entre dois nós na rede, e cada associação entre palavras aumenta em 1 o peso de sua aresta correspondente. Deste modo, quanto mais vezes uma associação aparece (na mesma ordem) no texto, maior é o peso dessa aresta na rede. As arestas obtidas são então representadas como uma matriz *N* x *N*, onde *N* é o número de vértices. A partir dessa modelagem, o texto passa a ser tratado simplesmente como uma rede complexa e é, portanto, submetido à análise sob essa perspectiva.

3. Medidas utilizadas

A representação de textos por redes complexas permite vários focos de análise através de suas diferentes medidas, e daí vem a nossa preocupação em encontrar medidas que melhor se ajustem às aplicações dessa área. Algumas dessas medidas foram implementadas no Núcleo Interinstitucional de Lingüística Computacional e utilizadas em experimentos [5]. Segue uma breve descrição de cada uma delas e como obtê-las:

- *Outdegree* (Grau): uma rede complexa é, por definição, um grafo direcionado com pesos nas arestas. Portanto, devemos levar em consideração ambos os graus de entrada e saída de cada nó. Entretanto, dado o método de formação da rede, cada aresta que incide em um nó tem uma aresta complementar que sai deste mesmo nó. Por isso os resultados para grau de saída e grau de entrada são sempre os mesmos, e com isso, apenas o grau de saída é levado em conta.
  O grau de cada nó *i* da rede é medido com a seguinte fórmula:

$$OD(i) = \sum_{j=1}^{N} W(j,i)$$

  Onde *i* e *j* são nós da rede e *W(j,i)* o peso da aresta *i*→*j*. A medida final é a média aritmética dos graus de todos os nós da rede.

- *Clustering Coefficient* (Coeficiente de aglomeração): para um nó *i*, o coeficiente é definido da seguinte maneira: quanto mais arestas existem entre um vizinho específico de *i* e os outros vizinhos de *i*, maior é o coeficiente de aglomeração. Isto é, o *CC* verifica quão próximo o subgrafo dos vizinhos do nó *i* da rede está

do grafo completo com o mesmo número de nós. O valor da medida é 1 quando todos os nós de um conjunto de vizinhos se conectam entre si. O valor para a rede toda é a média aritmética de todos os *CC*s. Matematicamente, temos: seja *R* o conjunto de todos os nós que se conectam ao nó *i*. Temos que $N_c = |R|$ e *B* é o número de arestas existentes entre os nós do conjunto *R*. O *CC* do nó *R* é calculado da seguinte maneira:

$$CC(i) = \frac{B}{\mathrm{N_c}\,(\mathrm{N_c} - 1)}$$

Se $N_c = 0$ ou $N_c = 1$, $CC(i) = 0$.

- *Components Dynamics Deviation* (Desvio na dinâmica dos componentes): o *CDD* é uma medida que indica a uniformidade com a qual novas palavras são inseridas em um texto. Seu cálculo é feito da seguinte maneira: inicialmente, têm-se todos os nós e nenhuma aresta. Então, cada aresta da rede é adicionada e o número de componentes conexos é contado (isto é, é verificado se um novo nó foi adicionado à rede parcial). O desvio da dinâmica dos componentes é o resultado da comparação entre a curva do número de componentes obtida pela adição de arestas e a reta ótima, na qual os nós são adicionados uniformemente a cada nova aresta. O valor da medida para toda a rede é a média aritmética do desvio para todos os pontos da curva. Sejam f_a função da curva supracitada, f_s a função da reta, *N* o número de nós na rede e *L* o número total de associações de palavras no texto. O *CDD* é medido da seguinte maneira:

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

- *Matching index* (Índice de Correlação): o *MI* é uma medida aplicada às arestas, e observa a similaridade entre a conectividade dos dois nós adjacentes à aresta em questão, isto é, quanto menor o valor do *MI*, menor é a semelhança entre as regiões do grafo conectadas pela aresta. O cálculo da medida é feito da seguinte maneira: o *MI* da aresta $i \rightarrow j$ (aqui representada por a_{ij}) é o número de conexões compartilhadas entre *i* e *j*, isto é, quantas conexões ambos têm em comum com um terceiro vértice *k*, dividido pelo número total de conexões entre *i*, *k* e *j*, *k* (exceto conexões entre *i* e *j*); como o grafo de um texto é direcionado, são tomadas como arestas compartilhadas apenas as quem possuem a mesma direção. Matematicamente, temos:

$$MI(i, j) = \frac{\sum_{k \neq i, j} a_{ik} \cdot a_{jk}}{\sum_{k \neq i} a_{ik} + \sum_{k \neq j} a_{jk}}$$

- *Hierarchical Degree* (Grau Hierárquico): para definir o *HD*, é necessário primeiramente definir os conceitos de distância e anel. A distância entre os nós *i* e *j* será tomada neste artigo como o número de arestas necessárias para se chegar do nó *i* ao nó *j*. Por exemplo, os vizinhos (diretos) de um nó sempre têm distância 1 do mesmo. Já o anel $R_d(i)$ de distância *d* do nó *i* é o conjunto de todos os nós que distam exatamente *d* de *i*. O grau hierárquico de distância *d* de um nó

i pode então ser definido como o número de arestas que conecta os anéis $R_d(i)$ e $R_{d-1}(i)$. Também é feita a diferenciação pela direção das arestas, resultando em graus hierárquicos de saída e de entrada. Nos experimentos que utilizaram o *HD*, a distância utilizada foi 2, com 8 variações: todas as combinações de entrada/saída, acumulado/não-acumulado, com peso/sem peso das arestas. O termo acumulado indica que, para calcular o grau hierárquico de distância *d*, os graus de distância 1 até *d*-1 devem ser somados ao valor final da medida.

- *Degree Correlation* (Correlação de Grau): A *DC* mede quão próximos estão os graus (foi utilizado apenas o grau de saída de cada nó) dos nós adjacentes a uma mesma aresta, averiguando assim quão uniforme a distribuição das arestas é na rede. Seu cálculo é feito da seguinte maneira: para cada aresta, os graus de seus dois nós adjacentes são obtidos, e então dois vetores podem ser formados, $X = (X_1, X_2, ..., X_n)$ e $Y = (Y_1, Y_2, ..., Y_n)$, onde *n* é o número de arestas na rede, X_i e Y_i são os graus dos nós de incidência e de partida da *i*-ésima aresta. O valor da correlação de grau utilizado neste artigo foi o coeficiente de Pearson entre os vetores *X* e *Y*, calculado pela seguinte fórmula:

$$DC(X,Y) = \frac{\sum_{t=1}^{n}(X_t - \overline{X}) \cdot (Y_t - \overline{Y})}{\sum_{t=1}^{n}(X_t - \overline{X})^2 \cdot \sum_{t=1}^{n}(Y_t - \overline{Y})^2}$$

Onde $\overline{X}$ e $\overline{Y}$ são as médias dos valores de *X* e *Y*. Os resultados da *DC* variam entre -1 e 1, e valores em módulo próximos de 1 indicam correlação linear.

A seguir, descrevemos o corpus utilizado em nosso experimento de autoria com redes complexas. Cada texto do córpus foi representado como indicado na Seção 2 e as medidas anteriores foram calculadas.

4. Experimento

O corpus deste experimento consistiu de 31 livros (disponíveis livremente na web) de 6 autores consagrados da língua inglesa. Os livros foram utilizados na sua versão original, e seus nomes e respectivos autores estão descritos na Tabela 1.

Tabela 1. Autores e livros pertencentes ao corpus deste experimento, com os títulos originais em inglês.

Autor	Livros
Charles Dickens	A Tale of Two Cities American Notes David Copperfield Great Expectations Hard Times Master Humphrey's Clock Oliver Twist The Old Curiosity Shop The Seven Poor Travelers
Ernest Hemingway	The Garden of Eden Green Hills of Africa
Lewis Carroll	Alice's Adventures in Wonderland

	Sylvie and Bruno Through the Looking-Glass
Pelham G. Wodehouse	My Man Jeeves Tales of St. Austin's The Adventures of Sally The Clicking of Cuthbert The Gem Collector The Man with Two Left Feet The Pothunters The Swoop The White Feather
Thomas Hardy	A Changed Man and Other Tales A Group of Noble Dames Desperate Remedies Far from the Madding Crowd The Hand of Ethelberta
Virginia Woolf	Jacob's Room Night and Day The Voyage Out

Este corpus, embora de tamanho limitado e contendo textos não muito longos, contém um conjunto de textos bastante distintos entre si (embora sejam do mesmo gênero), o que contribui para confiabilidade dos resultados obtidos.

Cada um dos textos sofreu a seguinte análise: as 2.000 primeiras associações de palavras do texto foram obtidas, e os valores de todas as medidas foram obtidas para cada segmento destes de maneira cumulativa. Com isso, cada par texto-medida gerou 2.000 valores, que foram organizados em curvas, como na Figura 4, onde a curva da esquerda representa os valores das medidas (a curva em preto é uma aproximação por polinômio de grau 3 obtido pelo Método dos Mínimos Quadrados Generalizado (*MMQG*)), e à direita tem-se a flutuação da medida com relação ao polinômio.

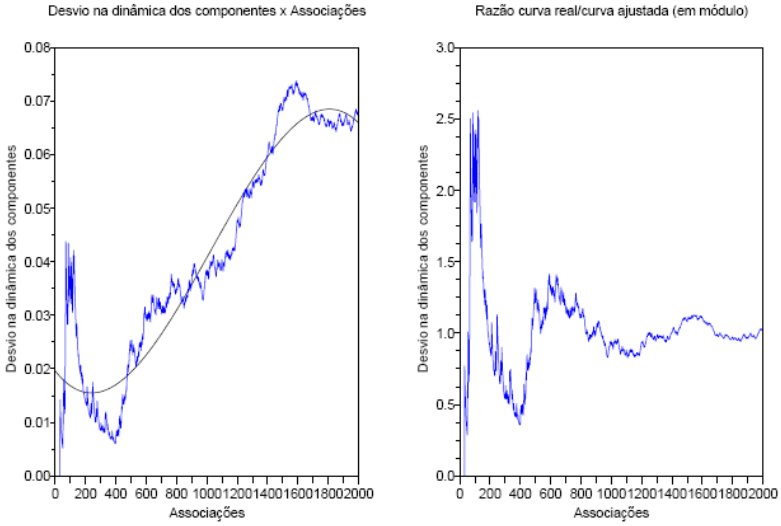


Figura 4. Gráfico das métricas da rede correspondente ao livro *American Notes*, de Dickens. À esquerda: cada valor de *n* no eixo das associações indica que as *n* primeiras associações foram utilizadas. À direita: gráfico da flutuação dos valores das medidas para textos de diferentes tamanhos.

Observando estas duas curvas é possível analisar a estabilidade de cada métrica, assim como é possível determinar um tamanho de texto para o qual a medida já se comporta de maneira uniforme, aumentando a confiabilidade do valor obtido.

Alguns exemplos de comportamento das medidas estão expostos nas Figuras 5-10, que correspondem ao livro *David Copperfield*, de Dickens. Pelas figuras é possível observar que na maioria dos casos os valores da flutuação diminuem conforme o tamanho do texto aumenta, eventualmente se aproximando de 1, onde o comportamento dos valores pode ser aproximado mais precisamente com o *MMQG*. A Figura 11 contém uma comparação entre os valores do *CDD* para 6 livros diferentes, com o intuito de demonstrar que os valores geralmente seguem um mesmo padrão de comportamento.

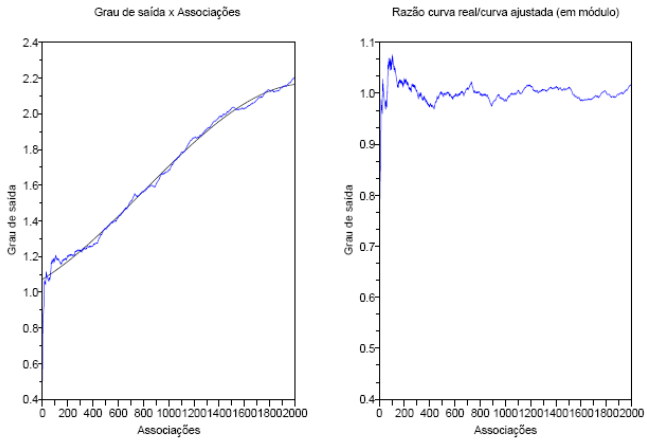


Figura 5. Gráficos para *OD*. É provavelmente a medida mais estável de todas, seguindo praticamente um comportamento monotonicamente crescente.

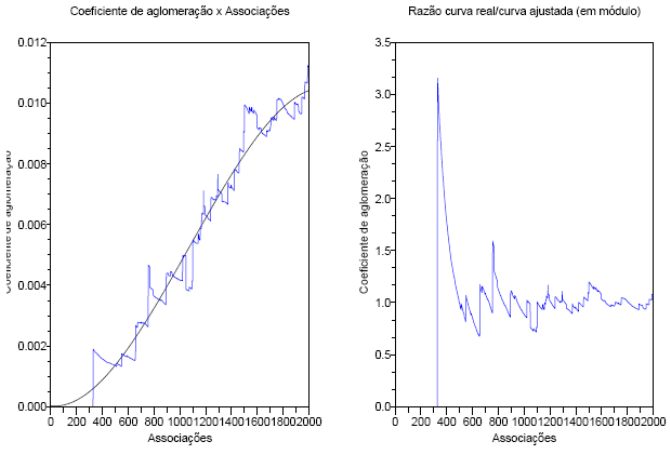


Figura 6. Gráficos para *CC*. Esta medida se caracteriza por somente crescer após um certo patamar de tamanho de textos.

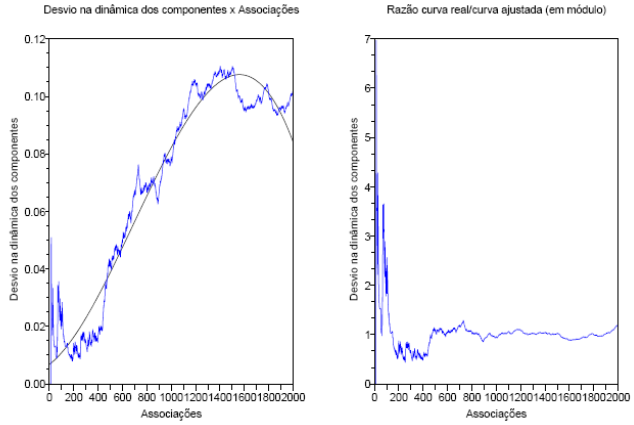


Figura 7. Gráficos para *CDD*. Para textos extremamente pequenos, esta medida é altamente instável, mas seu comportamento se altera para textos maiores.

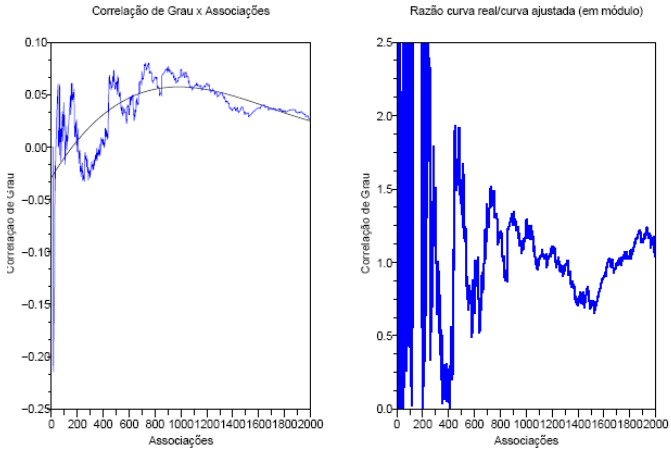


Figura 8. Gráficos para *DC*. Os valores desta medida tendem a ficar sempre próximos de 0, diminuindo assim a precisão da medição da flutuação.

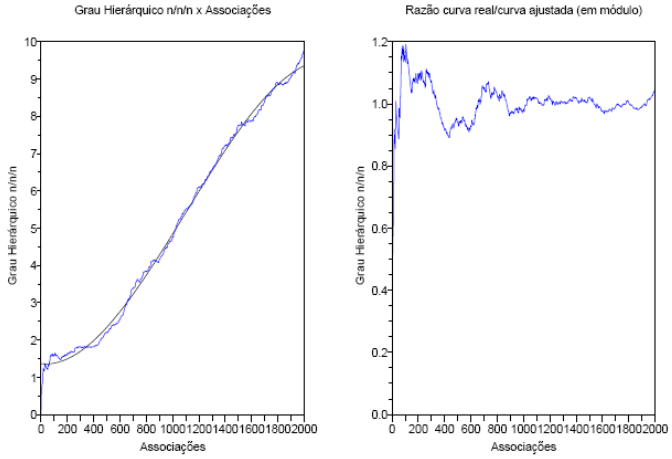


Figura 9. Gráficos para *HD*. Resultados bastante próximos aos do *OD*. Entre as diversas variações do *HD*, os resultados são geralmente muito próximos.

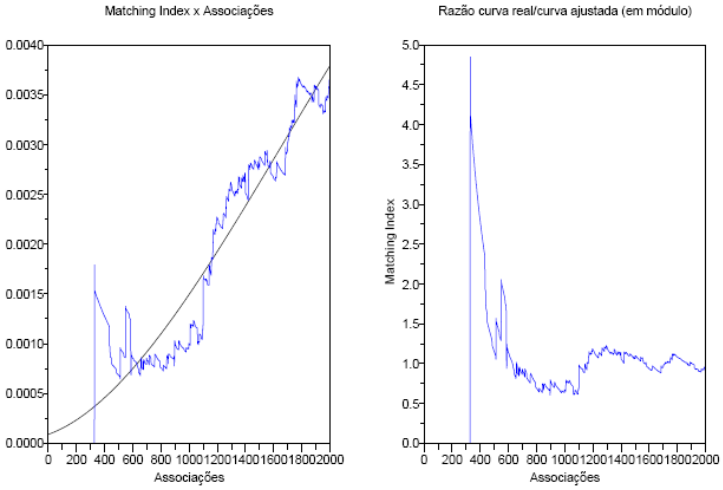


Figura 10. Gráficos para *MI*. Como é uma medida bastante parecida com *CC*, suas evoluções se comportam de modo bastante parecido.

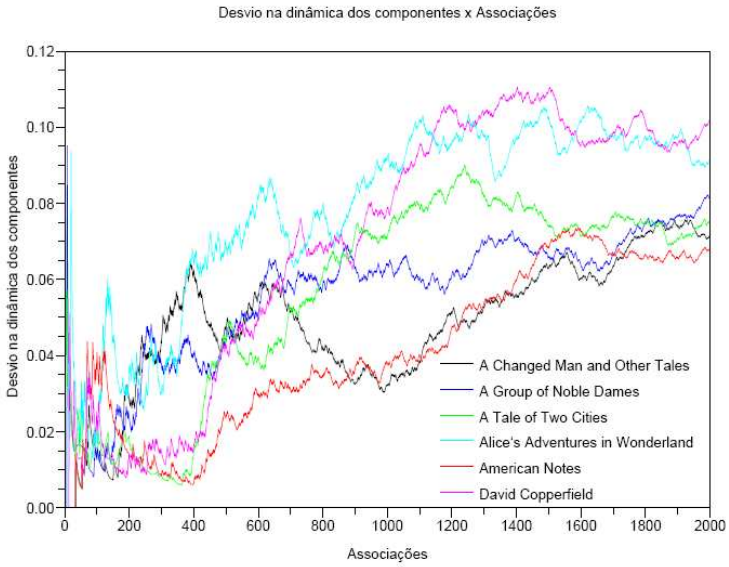


Figura 11. Comparação entre valores do *CDD* para vários livros diferentes.

Como pode ser visto nas Figuras 5-10, todas as medidas têm a mesma tendência de diminuir a flutuação conforme o tamanho do texto aumenta. Apesar disso, para o caso do *DC*, o tamanho de 2.000 palavras não é bastante para que os valores sejam previsíveis o bastante, e um destes casos de falha pode ser observado na Figura 12.

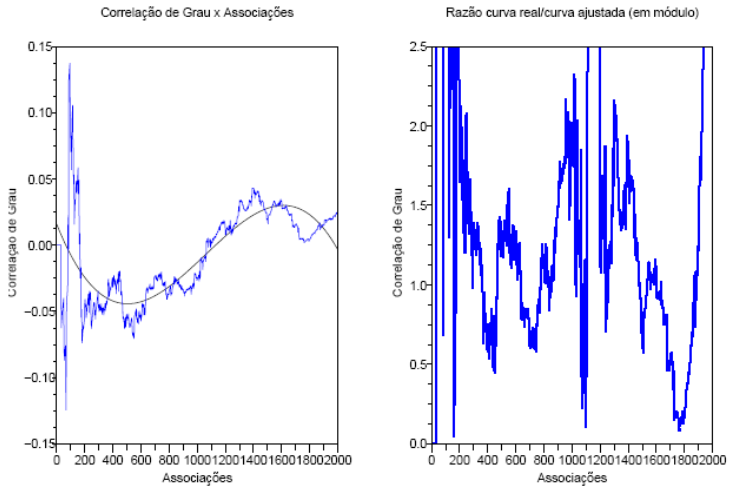


Figura 12. Exemplo de caso em que 2000 palavras não são suficientes para que a medida se estabilize.

Para todas as outras medidas, os resultados foram bastante positivos, com todos os textos seguindo aproximadamente o padrão definido pelos gráficos 2-7. É plausível assumir que os problemas encontrados com o *DC* provenham do fato do alcance de valores da medida ser muito próximo de 0, o que aumenta a flutuação relativa da curva (e pode implicar em valores menos estáveis).

Com estes resultados, é possível afirmar que todas as medidas estudadas, exceto o *DC*, são razoavelmente previsíveis, e portanto recomendáveis para utilização em futuros experimentos utilizando redes complexas. Além disso, foi possível se observar que textos com 2.000 palavras já têm tamanho suficiente para que as redes complexas tenham um bom desempenho com uma margem razoável de segurança, o que é um resultado bastante positivo, uma vez que os populares métodos estatísticos para problemas de PLN normalmente requerem uma quantidade muito maior de dados para fornecer resultados satisfatórios.

5. Conclusões

Os resultados obtidos ao longo deste trabalho comprovam que as diversas medidas para redes complexas estudadas, salvo a exceção da Correlação de Grau, são bastante previsíveis e podem, portanto, ser utilizadas em maior escala para fornecer resultados cada vez mais expressivos na área. Com isso, a possibilidade de se desenvolver sistemas de maior precisão usando redes complexas aumenta, permitindo assim avanços tanto no estudo das redes quanto na área de PLN.

Outro resultado importante é a observação de um tamanho ideal para aplicações de PLN envolvendo redes complexas, em torno de 2.000 palavras, que implica em textos bastante pequenos (poucas páginas), mas ainda assim fornecendo bons resultados. Esta informação é importante porque as aplicações de redes complexas normalmente são computacionalmente caras, exigindo boa quantidade de recursos e tempo e, como esperado, textos tão pequenos quanto possível colaboram para a construção de sistemas mais eficientes. Saber um tamanho de rede que seja estável e computacionalmente factível é de suma importância para o avanço da área.

Referências

[1] R. Albert and A. L. Barabási. Statistical mechanics of complex networks. Rev. Mod. Phys., 74:47–97, 2002.

[2] L. Antiqueira, M.G.V. Nunes, O.N. Oliveira Jr., and L. da F. Costa. Strong correlations between text quality and complex networks features. Physica A, 373:811–820, 2006.

[3] L. Antiqueira, T.A.S. Salgueiro, M.G.V. Nunes, and O.N. Oliveira Jr. Some issues on complex networks for author characterization. Revista Iberoamericana de IA, n. 36, 11:51–58, 2007.

[4] Lucas Antiqueira, Maria das Graças V. Nunes, Osvaldo N. de Oliveira Jr., and Luciano da F. Costa. Modelando textos como redes complexas. Technical report, Instituto de Ciências Matemáticas e de Computação, 2005.

[5] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. arXiv:cond- mat/0505185, 2006.

[6] R. Ferrer i Cancho and R. V. Sol´e. The small world of human language. P. Roy. Soc. Lond. B Bio., 268:2261, 2001.

[7] F. Peng, D. Schuurmans, V. Keselj, and S. Wang. Language independent authorship attribution using character level language models. In Proceedings of EACL, 2003.