UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE FÍSICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

MADĀYĀ DOS SANTOS FIGUEIREDO DE AGUIAR

REDES DE PALAVRAS EM TEXTOS ESCRITOS:
UMA ANÁLISE DA LINGUAGEM VERBAL UTILIZANDO REDES COMPLEXAS

Salvador - BA
2009

MADĀYĀ DOS SANTOS FIGUEIREDO DE AGUIAR

REDES DE PALAVRAS EM TEXTOS ESCRITOS:
UMA ANÁLISE DA LINGUAGEM VERBAL UTILIZANDO REDES COMPLEXAS

Dissertação apresentada ao Programa de Pós-Graduação em Física, da Universidade Federal da Bahia, como requisito parcial para a obtenção do título de Mestre em Física.

Orientador: Prof. Dr. José Garcia Vivas Miranda
Co-Orientador: Prof. Dr. Thierry Corrêa Petit Lobão

Salvador - BA
2009

Aguiar, Madaya dos Santos Figueiredo de
REDES DE PALAVRAS EM TEXTOS ESCRITOS: uma análise da linguagem verbal utilizando redes complexas / Madaya dos Santos Figueiredo de Aguiar. -- Salvador, 2009.
120 f. : il

Orientador: José Garcia Vivas Miranda.
Coorientador: Thierry Corrêa Petit Lobão. Dissertação (Mestrado - Programa de Pós-Graduação em Física) -- Universidade Federal da Bahia, Instituto de Física, 2009.

1. Rede Semântica. 2. Linguagem Escrita. 3. Força- Fidelidade. 4. Distância entre redes. I. Miranda, José Garcia Vivas. II. Lobão, Thierry Corrêa Petit . III. Título.

MADĀYĀ DOS SANTOS FIGUEIREDO DE AGUIAR

REDES DE PALAVRAS EM TEXTOS ESCRITOS:
UMA ANÁLISE DA LINGUAGEM VERBAL UTILIZANDO REDES COMPLEXAS

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Física, Instituto de Física, da Universidade Federal da Bahia.

Aprovada em 13 de agosto de 2009

Banca Examinadora

José Garcia Vivas Miranda – Orientador _____
Doutor em Ciencias Ambientais pela Universidad de La Coruña
La Coruña, Espanha
Universidade Federal da Bahia

Thierry Corrêa Petit Lobão – Co-Orientador _____
Doutor em Matemática pela Universidade de São Paulo
USP, Brasil
Universidade Federal da Bahia

Suani Tavares Rubim de Pinho _____
Doutora em Física pela Universidade de São Paulo
USP, Brasil
Universidade Federal da Bahia

Gilberto Corso _____
Doutor em Física pela Universidade Federal do Rio Grande do Sul
UFRN, Brasil
Universidade Federal do Rio Grande do Norte

À minha família, em especial minhas avós, Alaíde Baraúna e Wanda Figueiredo (*in memorian*), meus maiores símbolos de determinação, dedicação e AMOR.

AGRADECIMENTOS

Algumas pessoas dizem que agradecer é uma tarefa difícil... Talvez seja mesmo... E esta é a minha vez de, formalmente, lembrar daqueles que fizeram parte dessa caminhada.

Invariavelmente, devo começar por você, Zé Garcia... Que me aceitou como orientanda num momento uno da minha vida e teve a paciência, mesmo quando estava a perdendo e descabelado de preocupação, de respeitar meu tempo, minhas obrigações e meus estados tanto de felicidade quanto de desestímulo. Demorou... Mas hoje vejo você com outros olhos e consigo perceber o ser humano que é. Obrigada pela oportunidade, apoio e pelos puxões de orelha com um sorriso (essa é uma ótima técnica de fazer com que você se sinta cada vez mais "batráquia" e se mova, mesmo que como uma tartaruga).

Ao meu co-orientador, Prof. Dr. Thierry Lobão (e que para mim sempre será Tico). Você é peça singular não só durante o período do mestrado, mas da graduação também.

Aos meus professores do IF-UFBA, em especial a Profa. Dra. Suani Pinho, ao Prof. Dr. Roberto Andrade, a Profa. Dra. Maria Cristina Penido e a Profa. Dra. Jacira de Freitas por todos os momentos anteriores a este e, principalmente, pelo apoio, carinho e voto de confiança.

Aos amigos e colegas do IF-UFBA, FESC, CONES (especialmente a Nadja Maciel e Jaime Oliveira) e CECRF que me incentivaram sempre e contribuíram com idéias e sugestões. E, neste instante, permito-me falar em Gesiane Miranda Teixeira e deixar registrado que ela é irremediavelmente parte indissolúvel do processo. Serei grata por absolutamente tudo, viu?

Ao colaborador desse trabalho, Charles Novaes... Velhinho, você é... "massa!"

Aos amigos para toda uma vida: Alane Virgínia, Ana Carla e Andréia Bittencourt, Angelo Almeida, Flora Bacelar, Indianara Lima, Jéssica Araújo, Leonardo Bacelar, Mayane Nóbrega, Micael Oliveria, Milena Góes, Rejane Cristina, Renam Brandão, Saulo Cordeiro e Vanessa Romancini.

E por fim, aos meus maiores AMORES... "Meu bem", mainha, painho, minhas irmãs (todas elas), vovó, Tom-Tom, meus "únicos" tios e primos e os meus amigos-irmãos Maria Caroline Silva e Alan Santos... Que como toda "grande" família tem seus problemas, mas também tem muito AMOR para compartilhar. Saibam que vocês me tornam uma pessoa cada vez melhor e são, inegavelmente, o meu "Presente Precioso".

A todos vocês, meu eterno carinho!

Cheirinhos.

"O estudo da linguagem, ou mais especificamente, da linguagem humana, é o estudo da natureza" (RAMON FERRER I CANCHO, 2007).

RESUMO

Este trabalho avalia algumas características da linguagem escrita usando como base a teoria de redes complexas. O método está fundamentado nas idéias de rede semântica e no índice Força-Fidelidade introduzido por Teixeira (2007). Este índice, usado como parâmetro de construção da rede de palavras, revela as mais importantes associações entre palavras que ocorrem em textos orais ou escritos. Aqui, analisamos 50 textos literários escritos em 4 idiomas distintos (Inglês, Francês, Português (Brasil) e Espanhol). Primeiramente, cada um destes textos foi convertido automaticamente numa estrutura de rede semântica e, depois, um tratamento estatístico foi realizado com o fim de calcular os índices de redes complexas. Além disso, todos os textos passaram por um processo em que o comprimento das frases e o número de palavras foram mantidos inalterados, mas o vocabulário era escolhido aleatoriamente (texto embaralhado). Na primeira parte desta pesquisa, comparamos os índices de rede e a distribuição de graus para textos originais e embaralhados. Esta comparação sobre a organização da linguagem mostra diferenças quantitativas entre a rede de palavras dos textos originais e os textos embaralhados. É importante dizer que todas as redes dos textos originais apresentaram comportamento crítico em relação à Força-Fidelidade e características de redes livres de escala, enquanto que as redes dos textos embaralhados são redes aleatórias. Na segunda parte, selecionamos 36 textos, agrupamo-os em 3 classes (autor, conteúdo e idioma) e calculamos, dentro de cada classe, as distâncias Euclidianas entre pares de textos no espaço dos índices da rede. Depois, analisamos, usando Teste T, a diferença média destas distâncias entre os grupos formados por (a) textos com a mesma característica que define a classe e (b) outros textos da classe. Como resultado, mostramos que a estrutura topológica da rede crítica parece capturar a diferença entre textos de autores diferentes, embora não seja sensível a diferentes idiomas e conteúdos.

Palavras-Chave: Rede Semântica. Linguagem Escrita. Força-Fidelidade. Distância entre redes.

ABSTRACT

This work evaluates some characteristics of written language using the complex network theory as framework. The method is based on the ideas of semantic networks and on the Force-Fidelity index introduced by Teixeira (2007). This index is used as a parameter of words network construction and it reveals the most important associations among words that occur in oral or written texts. Here, we analyze 50 literary written texts in 4 different languages (English, French, Portuguese (Brazil) and Spanish). First, each of these texts was automatically converted in a semantic network structure and, after, a statistical treatment was done in order to calculate usual complex network indexes. Furthermore, all the texts were submitted through a process on which the length of sentences and number of words were kept unchanged, but the vocabulary was randomly chosen (random text). In the first part of this research, we compare the network indexes and the connectivities distributions to original and random texts. This comparison about organization of language shows quantitative differences between words network of original texts and the random texts. It is important to say that all the networks of original written texts presented critic behavior in relation to the Force-Fidelity and characteristics of scale-free networks, while the networks of random texts are random graphs. In the second part, we select 36 texts, group them in 3 classes (author, content and language) and calculate, inside each class, the Euclidian distances between pairs of text in the network index's space. After that, we analyze, using Test T, the average difference of these distances between the groups formed by (a) texts with the same characteristic that define the class and (b) others texts of the class. As a result, we show that the topological structure of critical network seems to capture the differences among texts of different authors, although it is not sensitive to different languages and contents.

Keywords: Semantic Network. Written Language. Force-Fidelity. Distance between Networks

LISTA DE FIGURAS

LISTA DE TABELA

LISTA DE QUADROS

SUMÁRIO

1. INTRODUÇÃO

> "Ao procurar explicar a linguagem, o homem está procurando explicar algo que lhe é próprio e que é parte necessária de seu mundo e da sua convivência com os outros seres humanos". (ORLANDI, 1999)

Desde o século IV a.C., o homem vem se dedicando a construir um sistema de escrita e analisar as palavras e seus significados (QUEIROZ, 2005). Porém, somente no início do século XX estes estudos ganham o status científico numa ciência que visa a descrever ou explicar a linguagem verbal humana — a Lingüística. Para esta, pouco interessa prescrever normas ou ditar regras de correção para o uso da linguagem (ORLANDI, 1999): a linguagem verbal, oral ou escrita, é o objeto de reflexão.

O homem produz a fala e a escrita com a utilização de signos (ORLANDI, 1999). É a partir deles que o homem se comunica, se identifica, cria uma representação ideal de mundo e expressa seus pensamentos e sentimentos. No caso da fala, em particular, Saussure constitui o signo lingüístico da combinação de dois termos: significante (imagem acústica[1] – imagem que se faz do som em nosso cérebro) e significado (conceito) (SAUSSURE, 2006).

Apesar do enfoque no significado destes signos dado pelos psicólogos, a linguagem não pode ser completamente entendida sem uma concepção adequada das formas pelas quais as palavras são evocadas e ordenadas nas frases (MANIS, 1973). Portanto, é a partir da organização do conjunto de unidades lingüísticas (as palavras), que se dá a linguagem verbal. Isto significa que apenas o entendimento de tais unidades (constituintes do sistema lingüístico) não é suficiente para que se compreenda a mensagem transmitida por uma frase (propriedade global). Assim, pode-se dizer que esta propriedade global emerge da interação entre as partes constituintes desse sistema. Deste caráter emergente, pode-se compreender a

[1] Para esclarecer a idéia de imagem acústica, segue uma citação do próprio autor: "O caráter psíquico de nossas imagens acústicas aparece claramente quando observamos nossa própria linguagem. Sem movermos os lábios nem a língua, podemos falar conosco ou recitar mentalmente um poema. [...] A imagem acústica é, por excelência, a representação natural da palavra enquanto fato de língua virtual, fora de toda realização pela fala" (SAUSSURE, 2006).

linguagem como um fenômeno complexo[2] e, portanto passível ao estudo do ponto de vista da Física Estatística.

Vinculada a esta área do conhecimento, utiliza-se a Teoria de Redes Complexas, ferramenta usada para explicar a dinâmica de sistemas complexos, para analisar um conjunto de textos literários com o fim de investigar algumas características da linguagem verbal humana. Vale ressaltar que os problemas abordados nesta dissertação referem-se a questionamentos já realizados por outros pesquisadores, como Ferrer i Cancho e Solé (2001, 2004), Caldeira (2005) entre outros, porém ainda não tinham sido tratados a partir do método proposto por Teixeira (2007).

1.1 O PROBLEMA DE PESQUISA

Como a linguagem está organizada? Ela apresenta uma característica própria que nos diferencia? Línguas diferentes apresentam os mesmos padrões que indicam uma característica da linguagem humana?

Segundo Bento (2004), é impossível desvincular o homem da linguagem. Ela é intrínseca à cultura humana e é uma forma de tomar consciência de nós mesmos. Para Gazzaniga *et al* (2006), é a "única entre as funções mentais em que apenas os seres humanos possuem um sistema verdadeiro[3] de linguagem". Talvez, exatamente por isso, a sua compreensão desperta tanto interesse de estudiosos das mais diversas áreas do conhecimento.

Suas duas principais formas de produção são a fala e a escrita (EYSENCK, 1994). Apesar de áreas cerebrais especializadas da linguagem terem sido reconhecidas há mais de um século (GAZZANIGA *et al*, 2006), o apelo quanto à localização não é suficiente para desprezar possíveis diferenças nos processos envolvidos nestas duas formas de produção.

De acordo com Eysenck (1994), a escrita e a fala apresentam semelhanças no fato de que parece haver um número de diferentes estágios envolvidos na sua produção.

[2] Uma condição necessária, porém não suficiente, para classificar-se um fenômeno como complexo é que ele ocorra em sistemas dinâmicos que estão fora do equilíbrio (PINHO, 1998).
[3] Embora o autor não tenha deixado claro o que ele chama por um "sistema verdadeiro de linguagem".

Considerando que na escrita, assim como na fala, existe um estágio de planejamento em que palavras são evocadas associativamente, de forma que se podem representar tais associações de palavras a partir de uma rede. Teixeira (2007) propõe um "índice", denominado Força-Fidelidade, sobre o qual é possível apresentar a "melhor" rede semântica de textos escritos ou orais. Essa rede é estabelecida a partir de certos critérios fundamentados pela Teoria de Redes Complexas.

Assim, este trabalho se apropria não só de um conjunto de programas desenvolvidos por Caldeira (2005) e Teixeira (2007), mas também do método proposto por esta última para a construção da rede de palavras, e examina alguns clássicos da literatura espanhola, francesa, inglesa e portuguesa (Brasil) buscando saber se as redes de palavras oriundas desses textos apresentam as mesmas estruturas topológicas quando comparadas com três classes específicas: autor, idioma e conteúdo. Além disso, foi investigado se estas estruturas se modificam quando se toma cada texto e o compara com um texto que é o seu correspondente aleatório. Ou seja, compara-se com o mesmo texto após ter sido submetido a um processo de embaralhamento.

1.2 OBJETIVOS

Esta pesquisa consiste de uma análise da linguagem verbal escrita utilizando como base de dados textos literários clássicos escritos em quatro idiomas distintos e como ferramenta de caracterização a Teoria de Redes Complexas. Com os conceitos advindos dessa teoria, buscou-se identificar o que representa a topologia da rede de associação das palavras, bem como aferir quantitativamente a organização que emerge deste processo mental dinâmico.

Assim, assumindo que a linguagem humana não se dá de maneira aleatória, presumiu-se que é possível extrair padrões de comportamento que podem ser identificados por características específicas ou universais da linguagem.

Para isso, foram estabelecidas duas proposições baseadas na idéia de agrupamento dos textos pertencentes à mesma classe no espaço dos índices da Rede Complexa. Tais

proposições estão fundamentadas em três critérios de agrupamento (conteúdo, idioma, autor) e na distância euclidiana entre pares de textos neste espaço dos índices.

1.3 ESTRUTURA DA DISSERTAÇÃO

Nos cinco capítulos que se sucedem, analisam-se, de maneira mais aprofundada, algumas das idéias expostas até aqui. Assim, no capítulo 2, faz-se uma discussão sobre a linguagem verbal escrita considerando tanto aspectos de localização cerebral quanto o processamento dela sob o ponto de vista da Ciência Cognitiva.

No capítulo 3, contextualiza-se a linguagem como um fenômeno complexo, apresenta-se uma revisão sobre a Teoria de Redes Complexas, um dos alicerces para este trabalho, e seu embasamento conceitual vindo da Teoria dos Grafos, além de expor o conceito de Força-Fidelidade proposto por Teixeira (2007) e uma brevíssima revisão de alguns trabalhos envolvendo tanto a linguagem humana quanto redes complexas.

No capítulo 4, aborda-se o método empregado neste trabalho de pesquisa desde a composição e justificativas da base de dados utilizada até o processo de tratamento dos dados e construção da rede de associação de palavras.

Nos capítulos 5 e 6, respectivamente, apresenta-se os resultados e discussões destes a partir da análise de alguns índices de caracterização/diferenciação usados pela Teoria de Redes Complexas e as considerações finais desse trabalho de pesquisa.

Por fim, seguem as referências e apêndices com informações relevantes a respeito do trabalho em questão.

2. UMA VISÃO SUCINTA SOBRE O PROCESSAMENTO DA LINGUAGEM VERBAL ESCRITA

No presente capítulo, apresenta-se uma pequena discussão sobre aspectos da linguagem verbal humana, tais como: características de uniformidade e localização cerebral, processamento da linguagem verbal escrita à luz da Ciência Cognitiva, léxico e memória, e um modelo de representação do conhecimento declarativo dos indivíduos (Rede Semântica).

2.1 A LINGUAGEM VERBAL HUMANA

> "É a linguagem, característica do humano, que descortina a possibilidade de não se agir/reagir mecanicamente a partir de estímulos discriminativos e, portanto, de se expressar a capacidade da intencionalidade". (LOFFREDO, 1999)

Definir linguagem não é uma tarefa simples. Talvez isso seja reflexo do caráter complexo que está vinculado ao entendimento desse fenômeno. Alguns estudiosos a conceituam a partir de suas funções, outros a vêem como uma das mais complexas características do cérebro humano (GAZZANIGA *et al*, 2006) visto que a fala e a linguagem simbólica (onde a escrita está inserida) são marcas unicamente da espécie humana.

Bordenave (1993), por exemplo, entende a linguagem como um código do processo de comunicação, uma representação do pensamento por meio de sinais que permitem a comunicação e a interação entre pessoas. Tal comunicação pode ser realizada através da linguagem verbal (que tem por unidade básica a palavra), não-verbal (expressa em gestos, imagens, sons entre outros) e mista (que se utiliza de ambas as formas anteriores).

Para Pinker (2002 *apud* PEREIRA, 2002), a linguagem, enquanto habilidade complexa e específica do ser humano, não é uma invenção cultural, e sim uma herança

biológica inata e universal inscrita no DNA da nossa espécie e que evoluiu através do tempo.

Os estudiosos do círculo de Bakthin constituem a linguagem como uma prática social, partilhada, uma entidade concreta e viva de signos ideológicos onde a palavra é "o modo mais puro e sensível da relação social" (BAKTHIN, 1979 *apud* PETRONI). Ou seja, eles não desvinculam a linguagem de sua natureza dialógica (FANTI, 2003).

Na Psicologia Cognitiva, um dos ramos da Ciência cognitiva, a linguagem é vista como um instrumento de pesquisa no qual, a partir da sua expressão, é possível acessar a memória do indivíduo através de um conjunto de signos lingüísticos — as palavras. Como foi dito no capítulo anterior, Saussure (2006) considera que este signo une um conceito a uma imagem acústica, e não uma coisa a uma palavra. Porém, o conceito deste signo não tem um significado único e geral. Ele é pessoal, pois sofre influência das idéias ou experiências vividas por cada indivíduo. De acordo com Fanti (2003), "a palavra aglutina o verbal e o não-verbal e constitui-se como enunciado, pois recebe acento de valor". Em outros termos, o significado de uma palavra não está na própria palavra, mas na mente de cada pessoa.

Considerando essas propriedades, a palavra, elemento de substancial importância para o embasamento do pensamento lingüístico, assume um dos papéis principais no trabalho em foco.

2.1.1 UNIFORMIDADE DA LINGUAGEM

Alguns pesquisadores, como Souza (2006), vêm se perguntando se o cérebro possui um sistema único para compreender e produzir qualquer idioma, ou se idiomas diferentes são processados de modos diferentes (KOLB e WHISHAW, 2002). Ou seja, línguas diferentes expressam organizações mentais diferentes?

A partir dos resultados de diversas pesquisas, parece que as semelhanças nos idiomas, mesmo que não sejam explicitamente aparentes, são muito mais fundamentais que as diferenças (KOLB e WHISHAW, 2002).

Chomsky e Pinker (*apud* KOLB e WHISHAW, 2002) argumentam que todos os idiomas têm características estruturais comuns, em virtude de uma base genética da

linguagem humana: os humanos apresentam uma capacidade inata de criar e usar a linguagem. Algumas evidências parecem favorecer essa hipótese (KOLB e WHISHAW, 2002), tais como:

- todas as pessoas em todos os lugares usam a linguagem, sendo sua complexidade dissociada da cultura de um grupo;
- a linguagem é aprendida na fase inicial da vida do indivíduo, entre 1 e 6 anos de idade, sem esforço aparente. Isso não significa que seu desenvolvimento não sofra influência da experiência vivida pelo indivíduo durante esse período. Caso não haja exposição a um idioma durante esse momento, suas habilidades de linguagem serão gravemente afetadas;
- todos os idiomas têm muitos elementos estruturais básicos em comum, possuindo regras gramaticais próprias que especificam como os vários termos da oração devem ser posicionados numa frase e como as palavras devem ser flexionadas de forma a transmitirem diferentes significados. Além disso, três classes gramaticais estão presentes em todos os idiomas, são elas: sujeito, verbo e objeto direto.

Essas evidências parecem indicar a existência de uma estrutura sintática preferencial que independe do idioma. Algumas pesquisas foram, e ainda são realizadas a fim de buscar mais informações sobre a natureza da relação entre linguagem e uma 'teoria da mente'. Uma dessas pesquisas foi realizada por Ashby e Bentivoglio (1993 *apud* ANTÔNIO, 2001) que investigaram diversas línguas (como sacapulteco[4], francês, espanhol, inglês, alemão, hebraico, quechua, rama, papago e japonês) onde essa estrutura preferencial se fazia presente.

2.1.2 LOCALIZAÇÃO DA LINGUAGEM

As pistas sobre uma região cerebral responsável pela linguagem surgiram no começo do século XIX quando neurologistas observaram pacientes com dificuldades de linguagem que apresentavam lesões no lobo frontal. Contudo, foi no final do século XIX e início do

[4] A título de informação, as línguas sacapulteco, quechua, rama e papago são ou foram faladas, respectivamente, pelo grupo maia da Guatemala, grupos andinos do território do Peru, por cerca de 30 pessoas de um grupo étnico da Nicarágua e indígenas da fronteira dos EUA e México.

século XX que se tornou claro que as funções da linguagem eram parcialmente localizadas, não apenas dentro, mas também em áreas específicas do hemisfério esquerdo (KOLB e WHISHAW, 2002). Isso só foi possível devido a uma mudança na observação médica pautada na investigação de cadáveres: a inacessibilidade ao espaço corpóreo levava os médicos a inferir, com base no que eles não podiam ver a causa do que podiam ver (FONSECA, 1998). A partir de então, ganham notoriedade estudos científicos referentes à afasia[5] e destacam-se estudiosos como Pierre Paul Broca (1824-1880), Karl Wernicke (1848-1905), Sigmund Freud (1856-1939), dentre outros.

Uma das primeiras propostas "localizacionista", discurso organicista onde se veiculava a idéia de que uma perturbação da linguagem corresponderia a uma área lesada e vice-versa, surge com o trabalho de Paul Broca. Em 1861, ele examinou o cérebro do cadáver de um homem que apenas pronunciava a palavra "tan" e fazia um juramento. O resultado deste exame indicava uma lesão recente no lobo frontal esquerdo. Com base neste e em outros casos, Broca concluiu que as funções da linguagem estavam localizadas no lobo frontal esquerdo. Esta região representada na Figura 1 é conhecida como Área de Broca.

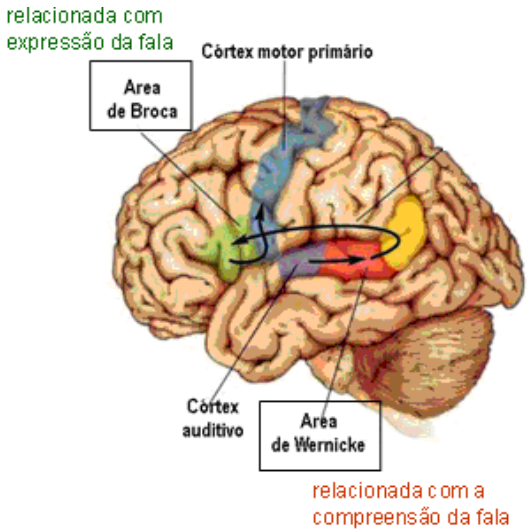


Figura 1. Localização da Área de Broca e Área de Wernicke
Fonte: http://www.freewebs.com/osnossospeterpan/etw5t6w.bmp

Nessa época, outros neurologistas acreditavam que a área de Broca poderia ser apenas uma região do hemisfério esquerdo que controlava a linguagem. Essa suspeita estava fortemente vinculada à possibilidade de uma relação entre a audição e a fala. A comprovação

[5] Afasia é um problema na linguagem causado por uma lesão cerebral (FONSECA, 1998).

deste fato veio com o resultado das pesquisas de Karl Wernicke. Ele avaliou pacientes que tinham dificuldades em compreender a linguagem após uma lesão na região posterior do lobo temporal esquerdo. Essa região, também indicada na Figura 1, é conhecida como Área de Wernicke.

Wernicke propôs um modelo de interação das duas áreas de linguagem do hemisfério esquerdo para a produção da fala: para falar palavras, mensagens são enviadas da área de Wernicke para a área de Broca por uma via que une essas duas áreas (o fascículo arqueado), assim área de Broca será acionada com um programa motor de produção de cada palavra que controla a articulação das palavras pelo aparelho vocal. Foram apenas após os estudos de Wilder Penfield, iniciados na década de 1930, que as áreas da linguagem do hemisfério esquerdo foram mapeadas de modo preciso e claro: a partir da estimulação elétrica foi possível identificar quatro regiões corticais importantes que controlavam a linguagem (as duas regiões clássicas mais a área suplementar da fala e as regiões faciais do córtex motor e somatossensorial).

Em **Sobre as Afasias**, Freud, ainda como neurologista, abala de forma radical a concepção localizacionista dos distúrbios da linguagem, criticando esse ponto de vista mecanicista do psiquismo. Ele propõe um circuito funcional da linguagem que apresentava uma relativa autonomia da topografia anatômica do sistema nervoso (LOFFREDO, 1999).

> [...] a cadeia dos processos fisiológicos no sistema nervoso não está em relação de causalidade com os processos psíquicos. Os processos fisiológicos não cessam mal se iniciam os psíquicos, pelo contrário, a cadeia fisiológica prossegue, só que, a partir de um certo momento, a cada elemento (ou a cada um dos elementos isoladamente) corresponde um fenômeno psíquico. O psíquico é assim um processo paralelo ao fisiológico (FREUD, 1979)

Com isto ele não quer negar que uma lesão cerebral produza efeitos no funcionamento lingüístico. Ele apenas não reduz a complexidade do lingüístico ao funcionamento cerebral (FONSECA, 1998).

Concordando com esta visão, Fonseca (1998) admite que tanto o funcionamento cerebral como o funcionamento da linguagem são realidades governadas por “leis próprias”, ou seja, uma realidade não se submete à lei de outro domínio: há autonomia, mas não há independência de domínios (cerebral e lingüísticos).

2.1.3 A LINGUAGEM VERBAL ESCRITA

A descoberta de regiões cerebrais específicas para o processamento da linguagem poderia nos levar a pensar que a produção da fala e da escrita ocorrem da mesma maneira. Entretanto, sabe-se que existem processos específicos relacionados a cada uma dessas importantes formas de linguagem. Atualmente, conhece-se ainda mais sobre a produção da fala do que da escrita. Para Eysenck (1994), isto pode ser justificado pelo tempo gasto no exercício da fala em relação à escrita e pelo papel que esta possui na sociedade.

Contrariando essa prática de pesquisa, este trabalho está voltado para a análise da escrita visto que por meio dela, a linguagem pode transcender às condições de tempo e espaço (QUEIROZ, 2005).

Do ponto de vista histórico (QUEIROZ, 2005), a humanidade viveu durante um longo período sem qualquer espécie de escrita, visto que esta pressupõe a existência da linguagem falada: a escrita teve origem apenas em meados do século IV a. C., com o surgimento do sistema de escrita *cuneiforme*.

Hoje, são considerados três diferentes sistemas de escrita (GAZZANIGA *et al*, 2006): o sistema alfabético (utilizado pela maioria das línguas ocidentais, no qual os símbolos aproximam-se dos fonemas), o sistema silábico (utilizado na escrita japonesa[6], onde cada símbolo reflete uma sílaba) e o sistema logográfico (no qual um símbolo único é utilizado para cada palavra ou morfema — o chinês é a língua que mais se aproxima desse sistema de escrita).

Do ponto de vista psicológico, ela é mais do que apenas um traço sobre o papel. Para Lacan (*apud* BENTO, 2004), a escrita, como linguagem, "é uma das formas do sujeito exercitar a sua subjetividade por meio da alteridade" (relação com o outro) sendo, portanto, uma das marcas do ser.

Do ponto de vista do processamento (EYSENCK, 1994), escrever é uma atividade de habilidade que envolve vários processos ou estágios diferentes. Uma das abordagens mais completas da escrita foi proposta por Hayes e Flower na década de 1980. De acordo com eles,

---

[6] A escrita japonesa constitui-se da associação de alguns milhares de caracteres chineses a dois sistemas silábicos: Hiragana e Katakana. http://www.invivo.fiocruz.br/cgi/cgilua.exe/sys/start.htm?infoid=915&sid=7

os processos-chaves na produção da escrita são: planejamento, geração da frase e revisão. A essência do que foi proposto é a seguinte:

- o planejamento envolve a produção de idéias e a sua organização em um plano de escrita que satisfaz aos objetivos do escritor. Estes planos de escrita apresentam forte dependência do conhecimento que o escritor possui sobre determinado assunto;
- o processo de geração das frases envolve a transformação do plano de escrita no ato de escrever;
- a revisão envolve a avaliação do que foi escrito, identificando deficiências para alterar o texto de tal maneira que ele se torne mais compreensível para o leitor.

Segundo Hayes e Flower (*apud* EYSENCK, 1994), existe uma seqüência natural de processamento da escrita que raramente ocorre, visto que tais processos parecem estar amarrados uns aos outros.

Apesar das diferenças existentes entre a fala e a escrita quanto à temporalidade, dependência espacial, velocidade de processamento, dentre outros, existem também semelhanças quanto ao número de diferentes estágios envolvidos na sua produção. Dos processos vinculados à escrita, é durante o estágio inicial de planejamento que os processos envolvidos na fala e na escrita são mais similares, com as diferenças aumentando aos poucos à medida que o processamento segue para o produto final.

Pesquisas realizadas por Taylor (1953 *apud* MANIS, 1973), Faigley e Witte (1983 *apud* EYSENCK, 1994), Kaufer *et al* (1986 *apud* EYSENCK, 1994) dentre outros, buscaram comparar estilos de escrita. O que se pôde observar é que:

- bons escritores mudavam freqüentemente a estrutura do plano de escrita à medida que novas idéias surgiam;
- escritores excelentes apresentaram uma média de, aproximadamente, 11 palavras por frase contra 7 palavras para escritores médios;
- bons escritores tendem a dar ênfase à estrutura e coerência dos argumentos expressos do que às palavras ou frases individuais;
- palavras familiares são mais utilizadas pois tornam o texto mais compreensível.

Em 1949, Zipf sugere uma lei que, em média, palavras curtas são mais utilizadas do que as longas. Sua justificativa encontra-se na idéia de mínima energia: quanto maior o

esforço envolvido na emissão ou escrita de uma determinada palavra, menor é a freqüência dessa palavra no linguajar cotidiano. Esta relação entre comprimento e freqüência da palavra foi, de fato, verificada no chinês, no latim e no inglês, sugerindo que esta é, provavelmente, uma característica de todas as línguas.

A preocupação, tanto do orador quanto do escritor, com a compreensão do texto está associada ao seu papel vital no que diz respeito à transmissão do significado. Segundo Miller (1954 *apud* MANIS, 1973), o significado total de uma frase é igual ao significado léxico (do dicionário) de suas palavras constituintes, acrescido do significado estrutural, transmitido através da ordem das palavras (regras sintáticas). Ou seja, a informação das palavras apenas não é suficiente para a compreensão da mensagem.

Esta organização sintática deve obedecer a uma regra que relaciona o comprimento da frase aos seus constituintes. Esta regra, conhecida como Lei Menzerath-Altmann (1954), determina que quanto maior for uma construção lingüística, menores devem ser seus constituintes (GRYZYBEK e KÖHLER, 2007). Esta lei também parece representar uma característica universal da linguagem verbal humana.

2.1.4 O LÉXICO E OS DOIS TIPOS DE MEMÓRIA

Segundo Gazzaniga *et al* (2006), o **léxico mental** é um estoque de informações sobre as palavras que inclui informação semântica (o significado da palavra), sintática (como as palavras são combinadas para formar uma frase) e os detalhes das formas das palavras (como são escritas e pronunciadas).

Nosso léxico mental contém dezenas de milhares de palavras onde, necessariamente, muitas delas se assemelham umas às outras. Tal dicionário dinâmico[7], embora contenha este estoque de informação, não deve ser organizado como um dicionário propriamente dito visto que esta não seria a forma de organização mais eficiente. Alguns modelos consideram que as seleções lexicais ocorrem a partir de um processo de competição e sobre influência do

[7] Esta característica está relacionada ao fato do léxico mental não possuir um conteúdo fixo: palavras podem ser esquecidas e novas palavras podem ser aprendidas.

contexto (GAZZANIGA *et al*, 2006). Isto é, as palavras não são processadas isoladamente, mas em um contexto entre outras palavras e por um processo de integração das palavras nas frases.

De maneira geral, é a partir deste conjunto de informações armazenadas no léxico que se pode externar o conhecimento humano através da linguagem. Então, pode-se dizer que a base deste conhecimento verbalizado é a **memória** (TEXEIRA, 2007).

Os psicólogos estudam a memória desde a metade do século XIX, e os psicólogos cognitivistas desenvolveram medidas sofisticadas de aprendizado e memória para pesquisas neuropsicológicas. Dessas medidas, se podem distinguir dois tipos de memórias nos seres humanos: memória implícita e memória explícita (KOLB e WHISHAW, 2002).

A memória implícita está relacionada à habilidade ou capacidade de demonstrar o conhecimento sem que, necessariamente, resgatem explicitamente as informações, e a explícita à capacidade de encontrar determinado objeto e indicar conscientemente que o objeto é correto. Essas classificações surgiram a fim de categorizar os diferentes processos de memória.

Esses dois tipos de memória estão relacionados a outros dois termos: a episódica (associada à memória do tipo explícita) e a semântica (relativa à memória do tipo implícita) (KOLB e WHISHAW, 2002).

Segundo Endel Tulving (1972 *apud* GAZZANIGA *et al*, 2006) a memória episódica está relacionada à memória para eventos, enquanto a memória semântica é aquela necessária para a compreensão e produção da linguagem e, portanto, está claramente conectada ao léxico mental.

Experimentos conduzidos por Nyberg *et al*. (1996, *apud* SCLIAR-CABRAL, 2002), utilizando tomografia de emissão de pósitrons (PET), demonstraram que o córtex pré-frontal esquerdo está mais envolvido na evocação da informação registrada na memória semântica do que o córtex pré-frontal direito. Lembre-se que neste hemisfério está localizada grande parte das regiões cerebrais responsáveis pela linguagem.

Inúmeros modelos vêm sendo sugeridos para explicar a estruturação da memória semântica. Um desses modelos representa graficamente o conhecimento declarativo a partir da verbalização de informações estruturadas em um conjunto de símbolos linguísticos interconectados. Esta estrutura é conhecida como **rede semântica**.

2.2 REDE SEMÂNTICA: UM RECORTE DA CIÊNCIA COGNITIVA

Segundo Gazzaniga *et al* (2006) “de modo geral, podemos dizer que as representações conceituais ou semânticas refletem nosso conhecimento do mundo real”

O surgimento da ciência cognitiva ocorreu no século XX e devido a várias razões diferentes. Uma delas foi a necessidade de tratar-se a cognição de uma maneira mais complexa do que a ciência até então tratava os fenômenos: entendendo-os como uma coleção de fatos.

Para a ciência cognitiva, a melhor metáfora da mente é a de um computador, e a cognição, uma forma de transformação de certos símbolos com procedimentos estabelecidos no interior desta máquina (QUEIROZ, 2000). Isto é, sendo cada símbolo uma unidade discreta, ele pode ser manipulado através de processos formais, com formatos e fórmulas específicas (DIAS, 2000).

Nesta nova área do conhecimento, suas principais abordagens são: psicologia cognitiva experimental, neuropsicologia cognitiva e a ciência cognitiva. Cada uma delas está envolvida num tipo específico de pesquisa.

Em particular, a psicologia cognitiva sugere modelos que representem a arquitetura cognitiva a partir do conhecimento declarado por indivíduos. Uma dessas formas de representação computacional ou matemática da estrutura cognitiva dos indivíduos é conhecida como rede associativa ou rede semântica.

As redes semânticas possuem as seguintes características (EYSENCK, 1994):

- os conceitos são representados por nós interligados para formar uma rede;
- estas interligações entre os nós podem ter graus de ativação (pesos) que relacionam, por exemplo, a conexão entre nós de uma mesma classe e nós de classes diferentes;
- a maior parte dos processos que ocorrem na rede serve para alterar os valores de ativação das interligações entre os nós;
- a forma pela qual a ativação se dissemina através da rede pode ser determinada por uma série de fatores. Alguns deles são: representação da ativação inicial, pela proximidade entre um nó e o ponto de ativação ou pelo intervalo de tempo que se

passou desde o início da ativação.

Um exemplo destes modelos foi proposto por Collins e Loftus em 1975 e está representado na Figura 2. Nela é possível notar as conexões e os graus de ativação relacionados à palavra cão e 3 palavras de outras classes: quanto maior o peso sobre a conexão, maior é a relação (força) entre um par de palavras. Isto pode ser um indicativo da relação entre palavras pertencentes à mesma classe e da relação.

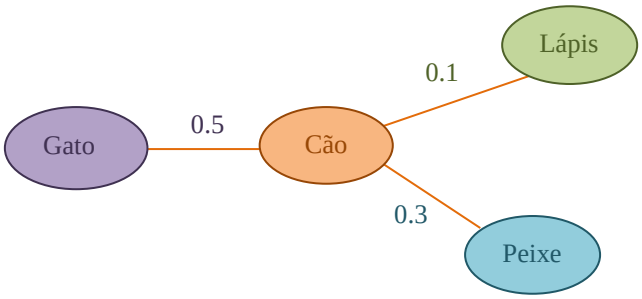


Figura 2. Diagrama esquemático de uma rede semântica simples com nós representados por conceitos e interligações entre estes nós indicando as diferentes analogias entre os conceitos.
Fonte: Adaptação de Eysenck (1994)

Apesar do conceito dessa rede de palavras ser fundamentado em idéias bastante simples, ela oferece um meio empírico de acesso à organização mental do conhecimento de tal maneira em que se pode transpor sua arquitetura organizacional e aplicar a Teoria de Redes Complexas ao estudo da Rede Semântica.

3. LINGUAGEM E COMPLEXIDADE

> [...] o cérebro humano tem sido apontado como um dos exemplos destes sistemas [sistemas complexos], visto que seu funcionamento global possibilita a realização de um conjunto de operações extremamente especializadas que cada neurônio jamais seria capaz de realizar individualmente e que tão pouco poderiam ser vistas de uma análise da dinâmica de interações de um neurônio com seu vizinho (PINHO, 1998).

O comportamento verbal humano tem fascinado estudiosos de várias áreas do conhecimento e tem sido estudado por meio de abordagens diversas. Mesmo com a nossa familiaridade com a palavra, existem regularidades estatísticas na linguagem que raramente se nota. Tais regularidades foram descobertas a partir de processos de contagem onde a comunicação (escrita ou falada) pôde ser analisada através da freqüência na qual ocorrem diferentes palavras. Esta relação entre palavras evocadas e sua correspondente freqüência de ocorrência se dá por uma Lei de Potência e foi observada por J. B. Estoup (1916) e, posteriormente, por G. K. Zipf (1949). Este tipo de característica sugere que a linguagem se comporta como um sistema complexo.

Apesar do conceito de complexidade ainda não possuir uma definição única, é possível classificar um sistema como complexo a partir de características que este sistema apresente. Algumas dessas características são (PINHO, 1998):

- possuir grande número de constituintes que interagem entre si e com o meio;
- exibir propriedades coletivas: o comportamento do todo não reproduz o comportamento das partes interagentes que o integram;
- evoluir de forma natural para um estado crítico através de um processo de dissipação de energia;
- podem apresentar um espectro de freqüência de eventos que obedece a uma Lei de Potência.

Na última década do século XX, físicos, psicólogos, sociólogos, biólogos, médicos, matemáticos e lingüistas[8] têm utilizado a Teoria de Redes Complexas na caracterização desses sistemas, bem como na descoberta de inúmeras relações imprevistas entre o funcionamento

[8] Ver, por exemplo, artigos referenciados em Newman (2003).

dos fenômenos humanos e o funcionamento de outros processos encontrados na natureza que, aparentemente, não tinham relação. Isso ocorre principalmente pela capacidade que esse método tem de representar "sistemas com comportamento dinâmico coletivo, rico e não-trivial" (PINHO, 1998).

Essa teoria, que tem origem na união entre a Física Estatística e a Teoria dos Grafos, trata o sistema como uma intricada rede de conexões entre pares de elementos denominados, respectivamente, como arestas e vértices ou nós da rede.

A linguagem humana pode ser vista como um fenômeno em que signos lingüísticos com significados próprios são organizados de forma a gerar uma estrutura com significado diferente da soma de cada unidade lingüística. Isto significa que é possível analisar o fenômeno lingüístico como um fenômeno complexo, a partir da emergência de uma propriedade global originada da compreensão de uma construção sintática (a frase) proveniente da interação e organização de unidades lingüísticas (as palavras). Da mesma forma que a compreensão de uma única frase de um texto não é capaz de refletir toda a mensagem deste texto.

A frase é a unidade básica de processamento lingüístico capaz de transmitir uma idéia. Segundo Caldeira (2005), ela é "a menor unidade para análise dos significados expressos nos textos, pois cada palavra isoladamente pode adquirir um significado que somente será identificado a partir do contexto".

Assim, este sistema de signos lingüísticos que surge de um processo mental dinâmico, complexo e associativo, pode ser modelado como uma rede complexa em que os vértices são representados pelas palavras evocadas e as arestas são as associações entre estas palavras. Dessa forma, as palavras que compõem cada frase formará um conjunto próprio em que cada palavra estará conectada a todas as outras palavras que constituem a frase. A este conjunto chamamos por clique. Se uma palavra for compartilhada por duas ou mais frases de um texto, então esta palavra conecta um clique a outro (Figura 3).

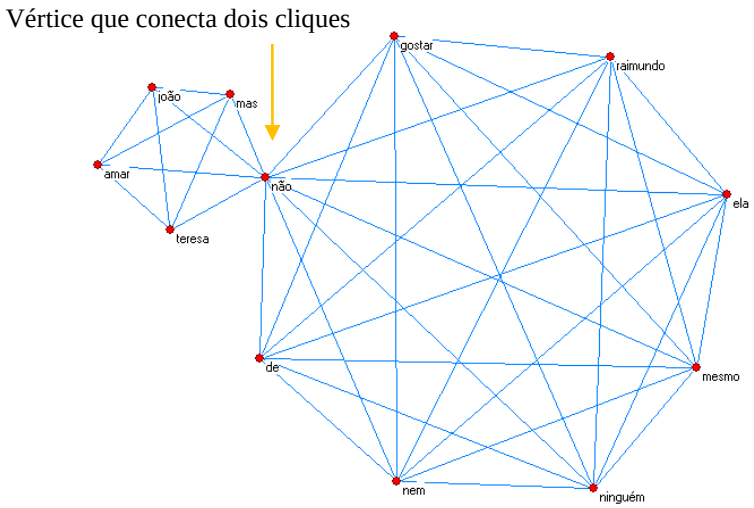


Figura 3. Ilustração da rede de palavras do texto '**João amava Teresa. Mas Teresa não amava João. Ela não gostava de ninguém, nem mesmo de Raimundo.**' destacando o vértice que conecta dois cliques distintos.
Fonte: Elaborado pela autora, 2009

Assumindo que, para cada par de palavras, existe um índice sobre a aresta que está associado a "força de interação" entre essas palavras, então a rede de palavras evocadas, seja numa comunicação oral seja escrita, pode ser dita como uma rede ponderada. Estes pesos refletem uma visão da análise do par de palavras sobre todos os pares que compõem o texto. Tal peso será chamado por Força-Fidelidade.

3.1 TEORIA DOS GRAFOS

A Teoria dos Grafos surge, de maneira incipiente, como fruto de um problema solucionado por Leonhard Euler em meados do século XVIII. Tal problema consiste em verificar a possibilidade de, partindo de um ponto qualquer de uma cidade chamada Königsberg, realizar um passeio completo atravessando, apenas uma vez, cada uma das sete pontes que cortam esta cidade. Note que este problema, conhecido atualmente como o Problema das Pontes de Königsberg, está vinculado à topologia desta cidade prussiana construída às margens do Rio Preguel.

Para modelar este sistema, Euler considerou cada massa de terra como um ponto e cada ponte como uma linha que conecta esses pontos conforme o mapa da cidade (Figura 4).

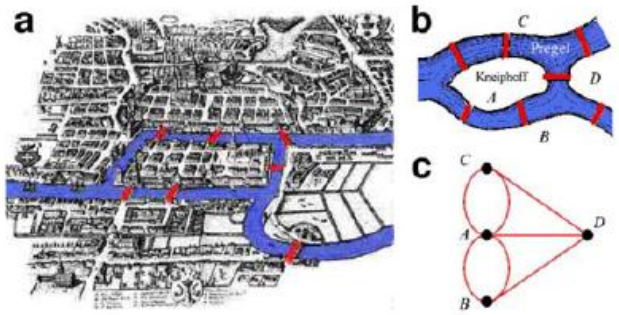


Figura 4. (a) Mapa da cidade de Königsberg, atual Kaliningrado (Rússia). (b) Representação esquemática das pontes de Königsberg com indicação de quatro massas de terra sendo uma delas correspondente à ilha Kneiphoff (A). (c) Ilustração do grafo que representa a cidade
Fonte: Amaral, 2004

Ele mostrou que é impossível executar tal passeio, visto que não poderia haver mais de duas massas de terra com um número ímpar de pontes. No caso de Königsberg, as quatro massas de terra estão conectadas por um número ímpar de pontes.

Apesar deste problema ter sido exposto como um questionamento local, esta análise pode ser realizada a qualquer rede de pontes para diferentes cidades. Além disso, a abstração proposta por ele tornou possível representar, topologicamente, um sistema a partir de um conjunto de pontos conectados por ligações. A esta estrutura denomina-se de Grafo.

Numa definição formal, um grafo G = (V, A) é uma estrutura composta por um par de conjuntos tal que V é um conjunto finito e não vazio constituído por elementos chamados vértices ou nós, e A é uma relação binária em V ($A \subseteq V \times V$). Cada um dos pares ordenados que constituam A é conhecido por arcos ou arestas[9]. Dessa forma, G pode ser visto como um conjunto de nós conectados entre si por arestas em que tais arestas não são paralelas.

Logo, seja G o grafo formado pelos conjuntos V={a, b, c, d, e} e A={(a,b), (b,c), (a,c), (b,d), (b,e)}, pode-se representá-lo conforme a Figura 5.

[9] O termo aresta é utilizado apenas em grafos não-direcionados (GALVÃO, 2006).

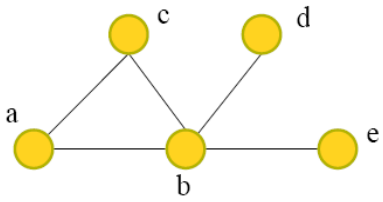


Figura 5. Exemplo de um grafo G composto por cinco vértices e cinco arestas
Fonte: Elaborado pela autora, 2009

Sabe-se que a cardinalidade do conjunto dos vértices e do conjunto das arestas determinam, respectivamente, a ordem (n) e o tamanho do grafo (m), enquanto o número de conexões (arestas) incidentes num vértice i, o grau (k) de i. Então, segundo a Figura 5, o tamanho e a ordem do grafo G é 5, e o grau do vértice b, por exemplo, é k = 4.

Observe ainda em relação à Figura 5, que G não apresenta qualquer conexão de um vértice i com ele mesmo, isto é, um elemento (i,i) no conjunto A. Portanto, G é classificado como um grafo sem laço.

Além disso, dois vértices quaisquer do grafo G estão conectados por apenas uma aresta. Ou seja, não existem arestas paralelas em G. Se G é sem laço e sem arestas paralelas, então G é um grafo simples.

Como existem muitos tipos de grafos e esta revisão não tem o objetivo de extinguir toda a discussão a respeito dessa teoria, apresentam-se apenas as classes que possuem relação com o objeto de análise desse estudo. Dessa forma, abordar-se, resumidamente, os grafos classificados como não-direcionado, ponderado, desconexo, sem arestas paralelas e sem laço.

(1) **Grafo não-direcionado**

Um grafo é dito como não-orientado ou não-direcionado quando não existe uma direção privilegiada que conecta os vértices do grafo, por exemplo, (a,b) = (b,a) (Figura 5). Caso a conexão entre os vértices tenha uma direção beneficiada, denomina-se de grafo direcionado ou orientado ou dígrafo (Figura 6).

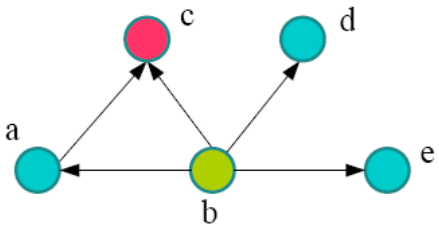


Figura 6. Exemplo de um grafo orientado (Dígrafo)
Fonte: Elaborado pela autora, 2009

Denomina-se por fonte um vértice pertencente a um dígrafo que apresenta grau de entrada 0 e grau de saída ≥1 e por sumidouro, aquele com grau de saída 0 e grau de entrada ≥1. Na Figura 6, b representa uma fonte e c um sumidouro.

(2) **Grafo ponderado**

Quando um valor numérico é atribuído às arestas de um grafo, conforme a Figura 7, designa-se este grafo de ponderado ou valorado.

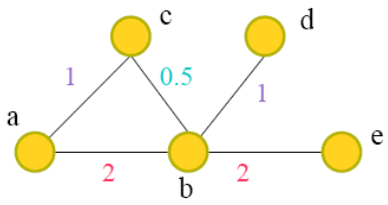


Figura 7. Exemplo de um grafo ponderado
Fonte: Elaborado pela autora, 2009

(3) **Grafo desconexo**

Um grafo é dito conexo se há pelo menos uma seqüência qualquer de arestas adjacentes que ligam um vértice qualquer a todos os outros vértices deste grafo. Se, pelo menos, um par de vértices não estiver ligado por alguma cadeia, como mostra Figura 8 (B), esse grafo é denominado desconexo.

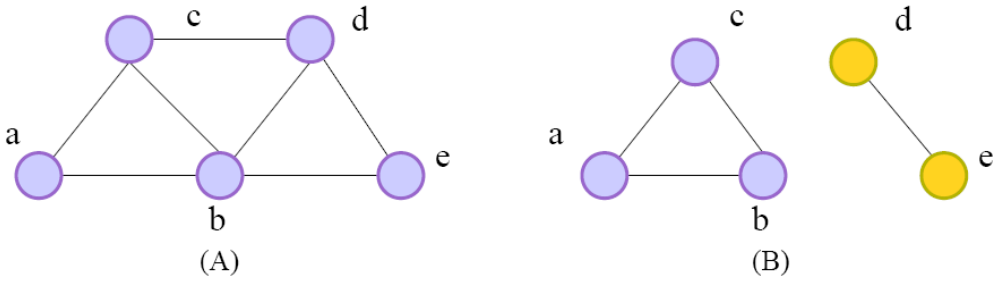


Figura 8. Representação de um grafo conexo (A) e outro desconexo (B)
Fonte: Elaborado pela autora, 2009

Chama-se de subgrafo de G um grafo G'= (V', A') tal que V'⊆ V e A'⊆ A como também A'⊆ V' x V' (Figura 9). Ou seja, como o conjunto de vértices e arestas que definem o subgrafo são subconjuntos daqueles que definem o grafo, um subgrafo é um subconjunto do grafo (GALVÃO, 2006). Se G' for um subgrafo completo[10] de G, como mostra a Figura 9, então G' forma um clique[11].

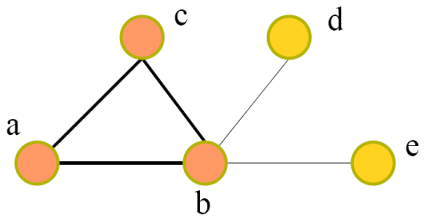


Figura 9. Clique formado pelo subgrafo *abc*
Fonte: Elaborado pela autora, 2009

Uma outra maneira de representar um grafo se dá pela construção de uma matriz onde seus elementos refletem a relação de vizinhança entre os vértices que compõem esta estrutura. Dessa forma, dois vértices são ditos vizinhos se eles compartilham a mesma aresta. Da idéia de vizinhança ou adjacência exposta acima, para um grafo de n vértices, se pode construir uma matriz quadrada M, denominada matriz de adjacência, cujos elementos M(i,j), assumem dois possíveis valores conforme a seguinte regra:

- se dois vértices i e j estão ligados por uma, e somente uma, aresta, M(i,j) = 1;

[10] Um grafo é dito completo quando existe uma aresta conectando cada par de vértices. Isso faz com que, num grafo de ordem n, cada vértice tenha grau k=(n-1). Na Figura 9, tem-se representado um clique de ordem 3 e grau 2.
[11] Neste trabalho, um clique será composto pelas palavras presentes em uma sentença e suas correspondentes associações.

- se não há uma ligação entre os dois vértices i e j, M(i,j) = 0.

Considerando o grafo simples G apresentado na Figura 5, pode-se construir uma matriz de adjacência M(5 x 5) e representá-la conforme a Figura 10.

	a	b	c	d	e
a	0	1	1	0	0
b	1	0	1	1	1
c	1	1	0	0	0
d	0	1	0	0	0
e	0	1	0	0	0

Figura 10. Matriz de adjacência M (5 x 5) relativa ao grafo simples G
Fonte: Elaborado pela autora, 2009

Assim, dado um grafo, é possível construir uma matriz de adjacência que o representa. Note que, com essa matriz, é possível visualizar apenas a relação de vizinhança do tipo primeiro vizinhos.

Para representar em uma única matriz todas as ordens de vizinhança entre todos os nós de um grafo, recomenda-se a utilização da matriz de vizinhança (ANDRADE et al., 2006). Por essa característica, a matriz de vizinhança acaba exibindo padrões de comportamento que dificilmente seriam percebidos explicitamente na matriz de adjacência. A Figura 11 mostra a matriz de vizinhança relacionada ao grafo G (Figura 5).

	a	b	c	d	e
a	0	1	1	2	2
b	1	0	1	1	1
c	1	1	0	2	2
d	2	1	2	0	2
e	2	1	2	2	0

Figura 11. Matriz de vizinhança M (5 x 5) relativa ao grafo G
Fonte: Elaborado pela autora, 2009

3.2 REDES COMPLEXAS

Uma rede é um conjunto de itens, chamados de vértices ou nós, com conexões entre eles (NEWMAN, 2003). A partir desta definição, pode-se dizer que uma rede é um grafo. Chama-se de Redes Complexas aquelas que apresentam um número muito grande de unidades que interagem de forma não-regular e que podem modelar, estatisticamente, sistemas dinâmicos, a partir de uma estrutura topológica. Utilizando as ferramentas vindas da Teoria de Redes Complexas, é possível (SANTANA, 2005):

- investigar sistemas, contendo milhões de elementos, macroscopicamente visto que o comportamento das partes não reproduz o comportamento do todo;
- encontrar e destacar propriedades estatísticas que caracterizam a estrutura e o comportamento de sistemas em rede;
- criar modelos de redes que ajudem a entender o significado dessas propriedades;
- prever o comportamento do sistema modelado em redes, baseado no comportamento das propriedades estatísticas.

Essa ferramenta de análise de sistemas reais multidimensionais, amplamente utilizada por muitos estudiosos das mais diversas áreas do conhecimento, oferece contribuições que consistem na estimativa de um conjunto de parâmetros das redes que revelam a sua topologia, grau de relacionamento, robustez, número de elementos, entre outros.

Na primeira década do século XXI, existia-se um conjunto composto por muitos parâmetros de análise de rede. Esta pesquisa se utiliza de alguns índices, dentre outros igualmente válidos. Estes índices são:

- tamanho e ordem da rede;
- grau de um vértice, grau médio e distribuição de graus;
- coeficiente de aglomeração e coeficiente de aglomeração médio;
- caminho mínimo de um vértice, caminho mínimo médio e diâmetro

3.2.1 ÍNDICES CARACTERÍSTICOS

Os índices característicos, listados na secção anterior, estão relacionados à métrica da rede. Suas definições são as seguintes:

(a) **Tamanho e ordem da rede**

A ordem (n) de uma rede, assim como de um grafo, corresponde ao número de vértices que a compõe, e o seu tamanho (m) ao número de arestas que ligam seus vértices.

(b) **Grau de um vértice, grau médio e distribuição de graus**

O grau k, também chamado de conectividade, de um vértice i de uma rede não-direcionada é determinado pelas conexões existentes entre esse vértice e seus primeiros vizinhos. Ou seja, é uma medida local que corresponde ao número de arestas incidentes no vértice (NEWMAN, 2003). Em contrapartida, o grau médio, <k>, é uma medida global da rede que se refere à média aritmética dos graus de cada vértice que compõe esta rede. Dessa forma, como nem todos os nós da rede têm o mesmo número de arestas, a distribuição de graus acaba nos informando qual a probabilidade P(k) de que um nó, aleatoriamente escolhido, tenha um número k de arestas. O histograma determinado por esta distribuição representa uma propriedade estatística fundamental na indicação da topologia da rede.

(c) **Coeficiente de aglomeração e coeficiente de aglomeração médio**

O coeficiente de aglomeração (C) de um vértice i é uma medida local da rede. Sua idéia pode ser apreendida considerando uma rede de amigos, por exemplo. Este índice representa a probabilidade dos meus amigos se conhecerem entre si. Em outras palavras, é a probabilidade de que os vizinhos de um dado vértice i serem vizinhos entre si. Outra forma de definir C é através do conceito de clique. Assim, C representa a probabilidade dos vizinhos do vértice i formarem um clique.

Na literatura, existem várias definições para o coeficiente de aglomeração (C) de um vértice i. Dentre elas, aquela apresentada por Albert e Barabási (2002) e expressa segundo a equação (3.1)

$$C_i = \frac{2m_i}{k_i(k_i - 1)} \tag{3.1}$$

em que

C_i é o coeficiente de aglomeração do vértice i

m_i é o número de arestas entre os vizinhos de i

k_i é o grau do vértice i

Então, se todos os vizinhos de i estiverem conectados uns aos outros este coeficiente valerá 1, e valerá 0 quando não estiverem conectados.

O coeficiente de aglomeração da rede é calculado através da média aritmética sobre todos os valores de C_i, isto é

$$C = \frac{1}{n}\sum_{1}^{n} C_i \qquad (3.2)$$

tal que

n = número de vértices da rede

C_i = coeficiente de aglomeração do vértice i

C = coeficiente de aglomeração médio (CAM)

(d) **Caminho mínimo de um vértice, caminho mínimo médio e diâmetro**

Como foi dito anteriormente, um caminho é conjunto de arestas adjacentes que conectam dois vértices quaisquer da rede. O menor caminho que ligue dois vértices quaisquer da rede é chamado de caminho mínimo (CM). A média dos caminhos mínimos entre um vértice i e os demais vértices da rede representa o caminho mínimo médio do vértice i. Portanto, quando se efetua uma média sobre todos os caminhos mínimos de todos os vértices que compõem a rede, tem-se o caminho mínimo médio (CMM) da rede complexa. O maior dentre os valores de caminho mínimo entre quaisquer dois vértices determina o diâmetro (D) da rede.

Com base neste conjunto de grandezas, podem-se identificar os tipos de redes. As principais, ou mais comuns, classes de redes são as redes regulares, redes aleatórias, redes livres de escala, redes de mundo pequeno, redes hierárquicas e redes modulares. Este trabalho fundamentou-se em três topologias:

- redes aleatórias;
- redes livres de escala (*scale-free*);
- redes de mundo pequeno (*small-world*)

3.2.2 TOPOLOGIA DE REDES

As **redes aleatórias** foram estudadas, exaustivamente e rigorosamente, pelos matemáticos húngaros Paul Erdös e Alfréd Rényi. De acordo com eles, se esta rede for composta por N vértices em que se conecta todo par de nós com uma probabilidade p, então, cria-se um grafo com, aproximadamente, pN(N-1)/2 arestas distribuídas aleatoriamente (ALBERT e BARABÁSI, 2002). Isto significa que não há critério algum que estabeleça privilégio na conexão de um vértice a outro.

Tais redes apresentam coeficiente de aglomeração médio igual a p, com $0 \leq p \leq 1$, e uma distribuição de graus característica do tipo normal (distribuição de graus de Poisson) com grau médio dado por p(N-1), para N infinitamente grande (Figura 12).

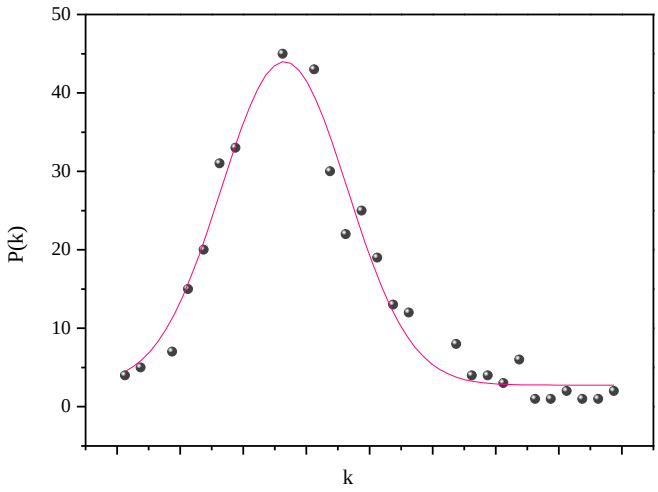


Figura 12. Representação da distribuição de graus de uma rede aleatória
Fonte: Elaborado pela autora, 2009

Apesar deste modelo ter sido aplicado apenas a fenômenos aleatórios, o interesse em uma variedade de sistemas tem levado os cientistas a considerar que tal modelo não apresenta princípios de organização que estão relacionados a uma diversidade de fenômenos reais (ALBERT e BARABÁSI, 2002). Ou seja, redes reais parecem não ser aleatórias (NEWMAN, 2003).

Tais redes reais, em geral, apresentam características de redes de mundo pequeno (*small-world*) ou livres de escala (*scale-free*). A **rede de mundo pequeno** surgiu do experimento realizado por um psicólogo social, Stanley Milgram, e seus colegas no final da

década de 1960. Este experimento teve a participação de diversas pessoas de cidades distintas dos Estados Unidos e mostrou que duas pessoas, que não tenham aparentemente qualquer relação, estão separadas por seis passos[12], em média. Isto se deve ao fato de que existe grande probabilidade de que essas duas pessoas tenham amigos que as aproximem. Da análise deste resultado, estava posta a noção de "mundo pequeno".

Em 1998, Watts e Strogatz propuseram uma modelo para descrever as redes de mundo pequeno reais. Tais redes não eram nem regulares[13] nem aleatórias (

Figura **13**). Elas apresentam coeficiente de aglomeração médio maior e caminho mínimo médio menor que uma rede aleatória de mesmo número de vértices e arestas, e a distribuição de graus pode se assemelhar a uma distribuição de Poisson devido a sua relativa homogeneidade.

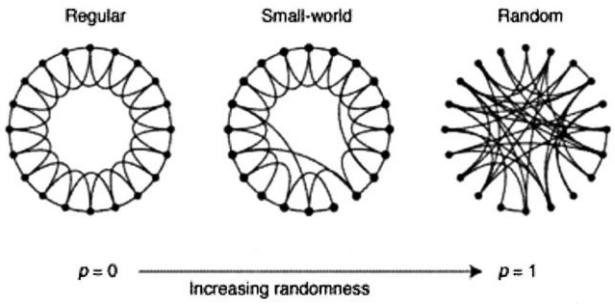


Figura 13. Representação de uma rede regular, mundo pequeno e aleatória composta por 20 vértices
Fonte: Albert e Barabási, 2002

Em geral, essas redes podem ser geradas de duas formas. Uma delas é promovida pela retirada de conexões de uma rede regular de grau k e posterior reconexão de vértices. Estas x reconexões aleatórias, com x = pNk/2, ocorrem com probabilidade p (Figura 14 (b)). Caso p=0, tem-se uma rede regular, e se p = 1, uma aleatória.

A outra maneira de gerar uma rede de mundo pequeno assume que todas as conexões existentes na rede regular são mantidas e novas ligações entre os vértices são realizadas aleatoriamente (Figura 14 (c)). Isso significa que, devido ao padrão da rede regular, a probabilidade de existirem vértices desconexos é nula.

[12] Um passo é o mesmo que uma aresta pertencente a um caminho.
[13] Redes regulares são aquelas em que $k_i = k$, $\forall i$.

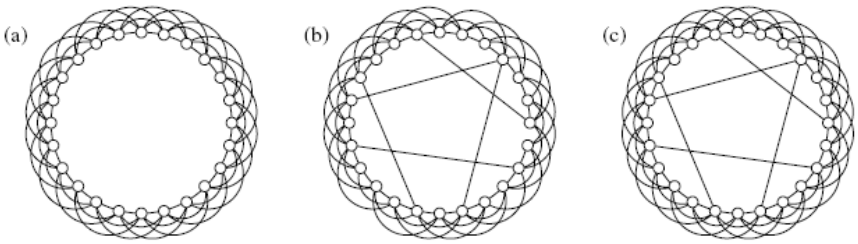


Figura 14. Representação de uma rede regular (a) e rede de mundo pequeno segundo o Modelo de Watts e Strogatz (b) e o Modelo de Newman e Strogatz
Fonte: Newman, 2003

Por fim, outro tipo de rede também muito freqüente na natureza são as **redes livres de escala** (Figura 15). Elas apresentam um arranjo de vértices e arestas que as tornam mais robustas a ataques aleatórios, porém mais vulneráveis a ataques dirigidos. Este arranjo está vinculado ao crescimento da rede. Um dos modelos mais usados para gerar estas redes é o proposto por Barabási e Albert em 1999 (ALBERT e BARABÁSI, 2002): partindo de um número pequeno de nós m_0, acrescenta-se um novo nó com $m \leq m_0$ arestas que conectam o novo nó a m diferentes nós já presentes nos sistemas.

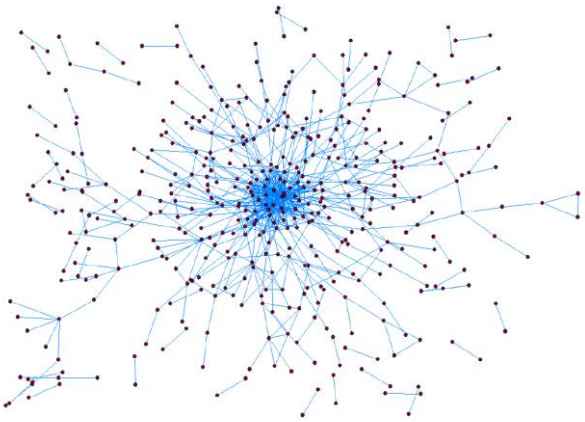


Figura 15. Exemplo de uma rede livre de escala
Fonte: Elaborado pela autora, 2009

Segundo este modelo, a probabilidade Π que o novo vértice conecte-se a um vértice i depende do seu grau k_i. Esta probabilidade é dada pela equação

$$\Pi(k_i) = \frac{k_i}{\sum_i k_i} \qquad (3.3)$$

Ou seja, essas ligações ocorrem de forma preferencial.

Depois de t passos de tempo, a rede é composta por N= m_0+ t nós e mt arestas. Dessa forma, quanto maior é o grau de um vértice, maior é a probabilidade de que ele receba mais vizinhos na iteração posterior. Nesse processo de crescimento surgem muito vértices com poucas ligações e poucos vértices com muitas ligações. A esses vértices de grau alto denominamos *hubs* ou concentradores.

Essas redes são, basicamente, identificadas por uma distribuição de graus na forma de lei de potência.

$$P(k) \sim k^{-\gamma} \tag{3.4}$$

no qual γ representa a inclinação da reta no gráfico log-log.

3.3 FORÇA-FIDELIDADE — UM POSSÍVEL ÍNDICE CARACTERÍSTICO PARA REDE CRÍTICA DE PALAVRAS

Em seu trabalho de mestrado, Teixeira (2007) criou o conceito de Força-Fidelidade que pode ser visto como a junção de dois conceitos: o de Força entre pares de palavras e o de Fidelidade.

A idéia de Força entre pares de palavras evocadas a partir da utilização da técnica de associação livre discreta[14] foi proposta por Nelson Douglas e colaboradores (NELSON *et al*, 1999 *apud* TEIXEIRA, 2007) e serviu como base para o trabalho sobre redes semânticas e redes complexas de Steyvers e Tenenbaum (2005).

Teixeira (2007) aplicou essa idéia de força entre pares de palavras para o discurso de indivíduos utilizando a técnica de livre associação[15]. Com o fim de estabelecer uma conexão entre estas duas idéias, ela propôs que estes dois conceitos fossem definidos a partir da

[14] Técnica onde o indivíduo associa, livremente, uma palavra à outra. Assim, dada uma palavra x, este indivíduo deve responder a primeira palavra y que lhe vier à mente.

[15] Com o fim de substituir a prática de hipnose, S. Freud introduz a técnica de livre associação. Neste método, a seqüência da comunicação deve seguir a fluidez do próprio pensamento, que vai surgindo de forma espontânea, sem que haja exigência de clareza, coerência, concisão, modo e relevância existentes em conversas cotidianas. http://www.scielo.br/pdf/agora/v5n2/v5n2a04.pdf

freqüência ou probabilidade de ocorrência de um par de palavras.

Dessa forma, tem-se que

(i) Força é a probabilidade de um par de palavras ocorrer em uma das frases do discurso.

(ii) Fidelidade é a probabilidade de um par de palavras sempre ocorrer nas frases que contêm ao menos uma dessas palavras.

Para melhor compreender estes conceitos, considere um conjunto finito C onde cada elemento desse conjunto é representado por uma frase e C_i como o subconjunto das frases em que a palavra i está presente.

Seja $S_i = |C_i|$ o número de elementos do subconjunto C_i, isto é, S_i representa a cardinalidade de C_i. Se tomar um par de palavras n e m quaisquer tal que $n, m \in C$, tem-se, respectivamente, que C_n e C_m são os subconjuntos formados pelas frases onde as palavras n e m ocorrem. Logo,

$$C_{n,m} \equiv C_n \bigcap C_m \tag{3.5}$$

é a intersecção entre os subconjuntos C_n e C_m e a cardinalidade deste conjunto será expressa por

$$S_{n,m} \equiv \left| C_n \bigcap C_m \right| \tag{3.6}$$

Ou seja, $S_{n,m}$ é o número de frases em que o par das palavras n e m ocorrem.

Pode-se representar o que foi exposto anteriormente através de um diagrama (Figura 16). Assim,

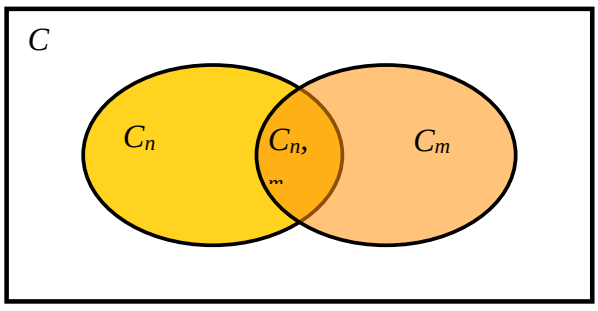


Figura 16. Diagrama do conjunto finito C
Fonte: Adaptação: Teixeira (2007)

Considerando as definições e o diagrama da Figura 16, podem-se escrever matematicamente os conceitos de Força e Fidelidade em função das cardinalidades dos conjuntos envolvidos. Logo, tem-se para a definição de força a expressão

$$F_O(n,m) \equiv \frac{\left|C_n \bigcap C_m\right|}{\left|\bigcup_{i=1}^{N_p} C_i\right|} = \frac{S_{n,m}}{N_S} \tag{3.7}$$

em que N_p é o número total de palavras e N_S o número total de frases do discurso.

E para a Fidelidade,

$$F(n,m) \equiv \frac{\left|C_n \bigcap C_m\right|}{\left|C_n \bigcup C_m\right|} = \frac{S_{n,m}}{S_n + S_m - S_{n,m}} \tag{3.8}$$

Dessas duas idéias, Teixeira (2007) define Força-Fidelidade como o produto destes dois conceitos e obtém a expressão

$$FF(n,m) = F_O(n,m) \cdot F(n,m) = \frac{\left|C_n \bigcap C_m\right|}{\left|\bigcup_{i=1}^{N_p} C_i\right|} \cdot \frac{\left|C_n \bigcap C_m\right|}{\left|C_n \bigcup C_m\right|} \tag{3.9}$$

Cada um destes índices pode assumir valores entre 0 e 1: será nulo quando as palavras nunca ocorrerem juntas ($|C_n \cap C_m| = 0$, isto é, $C_n \cap C_m = \varnothing$) e 1 quando todas as frases do discurso contiverem o par de palavras.

A diferença entre eles é que o primeiro nos indica a importância do par de palavras em todo o discurso, o segundo apenas nas frases em que este par ocorre e o terceiro considera ambas as situações.

Portanto, assume-se que a rede de associação de palavras advinda da linguagem oral ou escrita é uma rede ponderada onde os pesos sobre as arestas são os valores correspondentes à Força-Fidelidade dos pares de palavras que foram evocadas.

Note que, segundo a equação (3.7), (3.8) e, conseqüentemente, a (3.9), o tamanho do texto exerce grande influência sobre os valores da Força, Fidelidade e Força-Fidelidade. Buscando minimizar esse efeito, propõe-se que as equações (3.7) e (3.8) sejam rescritas.

Assim, tem-se que

$$F_{ON} = \frac{F_O - F_{O\min}}{F_{O\max} - F_{O\min}} \tag{3.10}$$

em que F_{Omin} e F_{Omax} são, respectivamente, o menor e maior valor de Força considerando todas as associações de pares de palavras que ocorreram no texto.

E

$$F_N = \frac{F - F_{\min}}{F_{\max} - F_{\min}} \tag{3.11}$$

para F_{min} e F_{max} o menor e maior valor de Fidelidade, respectivamente.

Essas equações estendem os valores de forma a preencher todo o intervalo [0,1]. Ou seja, não se atribui um peso demasiadamente alto a um nem insignificante a outro.

Das equações (3.10) e (3.11) tem-se

$$FF_N = F_{ON} \cdot F_N \tag{3.12}$$

A esta expressão chama-se de Força-Fidelidade Normalizada.

3.4 TRABALHOS ANTERIORES SOBRE LINGUAGEM UTILIZANDO REDES COMPLEXAS

Muitas pesquisas, nas mais diversas áreas do conhecimento, vêm utilizando a linguagem humana como objeto de estudo. Isto significa que existem inúmeras possibilidades de investigação sobre este fenômeno que têm sido analisado, principalmente, como um "bioproduto da interação social" (BOCCALETTI *et al*, 2006).

Dentre os estudos a respeito da linguagem como um fenômeno complexo, estão os trabalhos de Dorogovtsev e Mendes (2001), Ferrer i Cancho e Solé (2001, 2004), Steyvers e

Tenenbaum (2005), Caldeira (2005), Corso *et al*. (2006), Antiqueira *et al*. (2007), Teixeira (2007) e outros.

No artigo intitulado **Language as an evolving word web**, Dorogovtsev e Mendes (2001) consideram que a linguagem humana pode ser descrita como uma rede complexa não-direcionada em que cada vértice é uma palavra e as interações das palavras nas frases são representadas pelas arestas. Eles propuseram uma teoria estocástica de evolução da linguagem a partir de uma rede auto-organizada de interação entre palavras. Neste modelo, foi encontrada uma forma peculiar para distribuição de graus: duas regiões de leis de potência. Para eles, estes dois regimes emergem naturalmente não de regras da própria linguagem, mas da dinâmica de evolução da rede de palavras.

Com diversos trabalhos usando redes complexas para modelar a linguagem, Ferrer i Cancho e Solé (2001) utilizaram um conjunto de textos disponível no *British National Corpus*, sendo que os nós dessa rede representam as palavras, e suas arestas conectam palavras que aparecem no corpus pelo menos uma vez, em seqüência ou separadas por uma palavra. Dessa análise, foi mostrado que as redes de palavras apresentam características de redes de mundo pequeno (*small-world*) e livres de escala (*scale-free*). Posteriormente, Ferrer i Cancho *et al*. (2004) mapearam padrões sintáticos em redes de palavras em três idiomas distintos: alemão, checo e romeno. Deste estudo, observou-se que tais línguas partilhavam de padrões estatísticos não triviais, tais como características de mundo pequeno e distribuição de graus em escala, conseqüentes não da estrutura da frase (organização sintática) e sim de uma característica em escala global.

Em 2005, Mark Steyvers e Joshua Tenenbaum publicam um artigo com o resultado da análise de três tipos de redes semânticas: *WordNet*, *Roget's Thesaurus* e Associações de palavras. O primeiro desses tipos de redes foi inspirado em uma teoria psicolinguística e gerado a partir da conexão entre uma ou mais das 120 mil palavras e um ou mais dos seus 99 mil significados, ou seja, conceitos. Esta base de dados foi desenvolvida por George Miller e colaboradores (FELLBAUM, 1998; MILLER, 1995 *apud* STEYVERS E TENENBAUM, 2005).

O segundo tipo representa o produto do trabalho da vida do Dr. Peter Mark Roget (ROGET, 1911 *apud* STEYVERS E TENENBAUM, 2005) que classificou 29 mil palavras em uma mil categorias semânticas. Neste caso, a rede é direcionada, bipartida, composta por dois tipos de nós (palavra e categoria semântica) e construída de forma que cada nó está

conectado à sua categoria semântica (Figura 17).

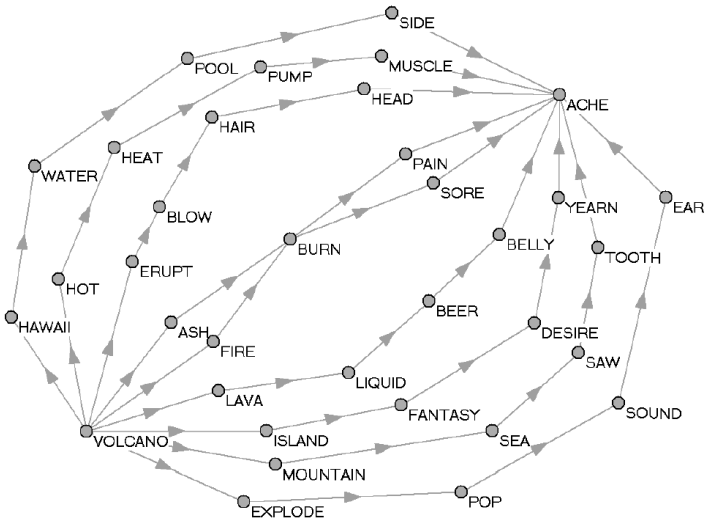


Figura 17. Parte da rede semântica direcionada formada por livre associação. Cada aresta ilustra uma associação entre a palavra sugestão e a resposta
Fonte: Steyvers e Tenenbaum (2005)

O último tipo de rede semântica avaliada por Steyvers e Tenenbaum (2005) foi construído a partir do banco de dados lingüísticos compilado por Nelson *et al* (1999, *apud* TEIXEIRA, 2007). Tal banco de dados constitui-se das associações livres discretas de 6 mil indivíduos considerando cerca de 5 mil palavras sugestões. Cada participante recebeu um conjunto de palavras sugestões e, para cada palavra, deveriam escrever a primeira palavra resposta que viesse em sua mente. Foram criadas duas redes: uma direcionada e outra não-direcionada. Na rede direcionada, dois nós *x* e *y* eram conectados por um arco se a palavra resposta *y* foi evocada da palavra sugestão *x* por pelo menos dois dos participantes. Na rede não-direcionada, os nós eram conectados se as palavras fossem relacionadas sem uma direção associativa. Por exemplo, quando a palavra HAVAÍ aparecia como palavra sugestão, a palavra resposta era FÉRIAS e vice e versa. Apesar da rede direcionada representar uma forma mais natural de associação, a rede não-direcionada possibilitou análise comparativa com redes tipo *small-word* e *scale-free*.

A Tabela 1 apresenta um sumário estatístico dos índices usuais para classificação da rede complexa considerando a rede de associação de palavras não-direcionada.

ÍNDICES	VALORES PARA A REDE DE ASSOCIAÇÃO DE PALAVRAS NÃO-DIRECIONADA
n	5018
<k>	22.00
CMM	3.04
D	5
CAM	0.19
γ	3.01

Tabela 1. Sumário estatístico dos índices usuais para classificação da rede complexa considerando a rede semântica não-direcionada
Fonte: Adaptado de Teixeira (2007)

Baseando-se na proposta do aparelho psíquico de Freud, Caldeira (2005) caracteriza a topologia da rede de conexões entre as palavras de textos escritos em dois idiomas distintos (inglês e português). Esta rede foi construída considerando cada frase como um clique e cada palavra compartilhada por duas frases diferentes tem a função de conectar estes dois cliques, formando uma rede complexa. A análise dos índices de tais redes e a distribuição de graus sugerem redes com características de mundo pequeno e livres de escala, assim como a análise feita por Ferrer i Cancho e Solé (2001), com expoente para a lei de potência ($P(k) \sim k^{-\gamma}$) de, aproximadamente, 1.6.

Corso *et al* (2006) utilizaram um conjunto de palavras evocadas por indivíduos de uma população para definir os vértices de uma rede e as conexões entre esses vértices foram estabelecidas pelos próprios indivíduos. O grafo resultante foi chamado de Rede de Palavras Evocadas (*Evoked Words Network*, EWN). Nesta pesquisa, foram consideradas três palavras *prime* (boca, doença e saúde) e dois grupos de indivíduos (pertencentes a um bairro de classe média e a um bairro pobre de Natal, Rio Grande do Norte - Brasil). Notou-se que as palavras evocadas entre esses dois grupos são bastante diferentes e refletem o status, escolaridade, hábitos de lazer e formas de expressão. A Figura 18 representa uma EWN de uma pessoa de classe média tendo como tema a palavra 'boca'. A distribuição de conectividades de todas as seis EWNs analisadas segue uma lei de potência com expoente $1.11 < \rho < 2.01$ o que indica uma estrutura *scale-free*[16].

[16] Como cada EWN não era suficientemente grande, n~100, o que implica numa estatística pobre, foi utilizada uma soma cumulativa, de tal forma que $\Phi(k) \sim k^{\rho}$ onde $\rho = \gamma + 1$.

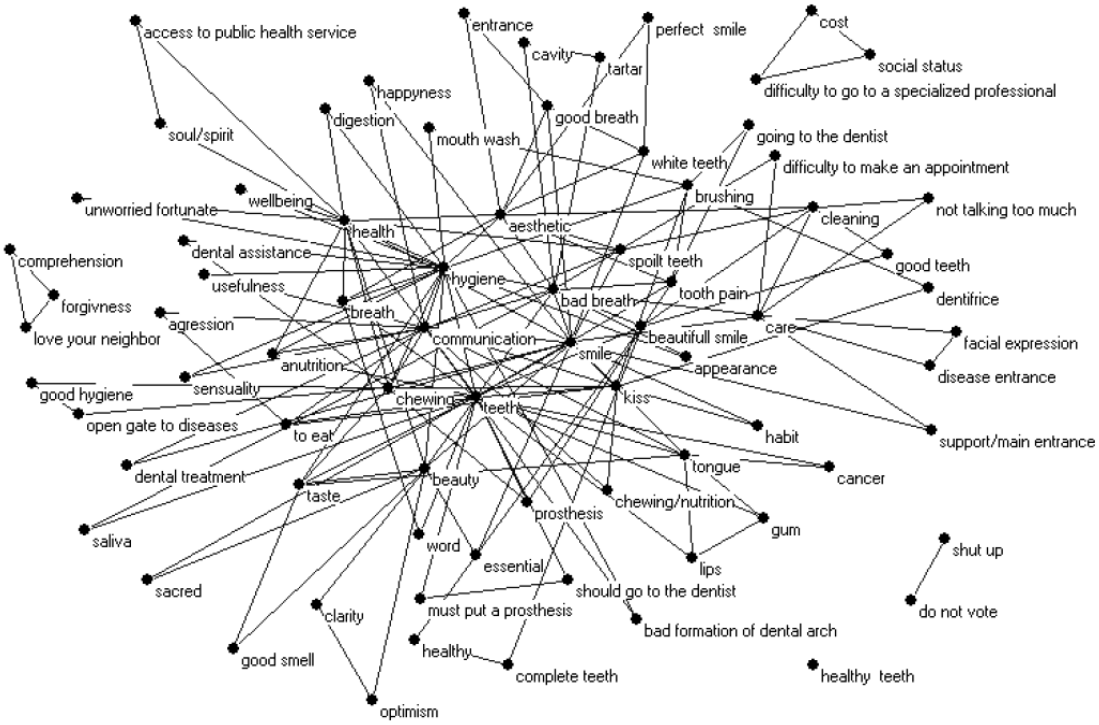


Figura 18. Esboço de uma EWN empregada por uma mesma pessoa que reside num bairro de classe média tendo como tema a palavra 'boca'
Fonte: Corso *et al* (2006)

Outro trabalho que utiliza redes complexas como ferramenta para análise e modelagem de textos foi o de Antiqueira *et al* (2007). Seus experimentos indicam que os parâmetros das redes complexas apresentam uma forte correlação com a qualidade de textos. Isto é, tais parâmetros podem ser utilizados para distinguir textos bons e ruins. A classificação dos textos dentro dessas categorias foi fornecida através de avaliação humana. Segundo seus resultados, o caminho mínimo é alto para textos com qualidade ruim, o que elucida a dificuldade dos escritores inexperientes em estabelecer conexões entre numerosos conceitos. Nos textos enquadrados como de boa qualidade, caminhos mínimos maiores foram observados. Entretanto, esse fato não estava correlacionado negativamente com as notas, ao contrário do que aconteceu com os textos da outra categoria. Esta modelagem, além de representar uma boa alternativa aos métodos de análise de texto, é independente da língua.

No ano anterior, Antiqueira *et al* publicaram um artigo intitulado **Some issues on complex networks for author characterization**. Neste trabalho, livros de 8 autores diferentes foram pré-processados de forma a obterem palavras como nós da rede e arestas que apontam para suas palavras adjacentes. Ou seja, a primeira palavra está conectada à sua

vizinha a partir de uma aresta direcionada e assim por diante. É importante destacar que não foi considerado os limites das frases e dos parágrafos determinados pela pontuação. Como resultados, obteve-se que as redes produzidas por cada um dos autores exibem características especificas, indicando que essas redes de palavras podem capturar características autorais.

Em sua pesquisa, Teixeira (2007) propôs outra forma de construção da rede semântica oriunda de 12 discursos coletados e transcritos, na íntegra, considerando a técnica de associação livre contínua estimulada por um tema (o "Eu"). A rede de associação das palavras foi gerada a partir do conceito de Força-Fidelidade (*FF*). Este parâmetro de filtragem, discutido na secção anterior, representa um índice para determinação da rede crítica, isto é, a rede que melhor representa a estrutura de associação entre conceitos do discurso. Esta rede é obtida através de um ajuste no valor da Força-Fidelidade em que vértices isolados e arestas cujos pesos sejam menores do que um valor FF_i são retirados da rede. Dessa forma, à medida que a Força-Fidelidade aumenta, permanecem na rede apenas as associações com valores maiores ou iguais a FF_i.

Neste processo de filtragem, notou-se que a representação gráfica do diâmetro em função da Força-Fidelidade apresentava um ponto de máximo valor. A este ponto chamou-se de ponto crítico e à Força-Fidelidade correspondente de Força-Fidelidade Crítica (*FFc*).

As topologias dessas redes apresentaram características de rede de mundo pequeno e indicações de redes modulares e livres de escala[17]. Além disso, através da análise das palavras presentes nos núcleos da rede, observou-se que tais redes se organizam de maneira categorizada. A Figura 19 ilustra a rede de associação das palavras evocadas pelo indivíduo I2 para o valor de Força-Fidelidade Crítica (*FFc*).

[17] Segundo Teixeira (2007), a média do vocabulário utilizado em todo o discurso foi de 638 palavras. Com a filtragem, o número de vértices decai fazendo com que, para a Força-Fidelidade Crítica, este valor seja ainda menor. Portanto, não é possível, considerando a distribuição de graus, uma classificação precisa desta rede crítica como uma rede livre de escala.

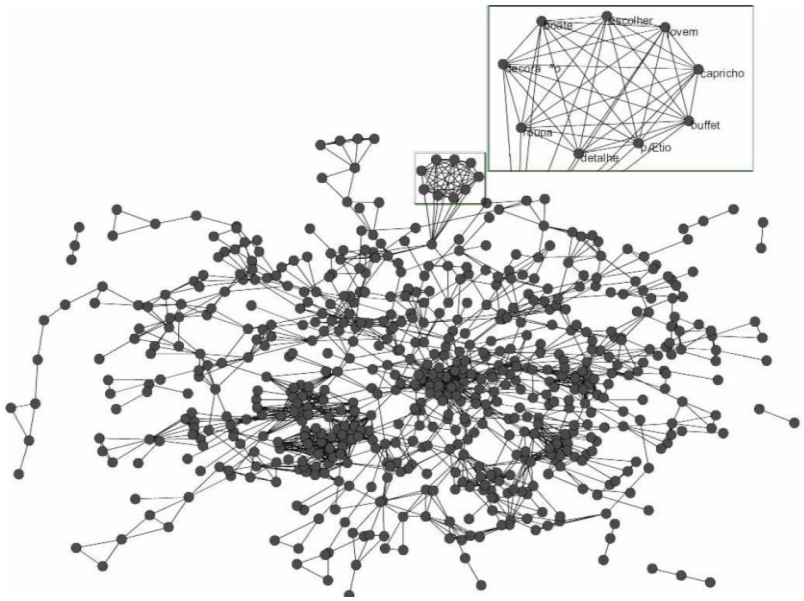


Figura 19. Rede crítica para o discurso do indivíduo I2 com detalhe de uma subrede
Fonte: Teixeira (2007)

De acordo com Teixeira (2007), apesar de existir certa variabilidade no valor de *FFc* para os distintos discursos analisados, todos apresentaram pontos de máximo diâmetro, com uma rede crítica bem definida e com topologias similares, indicando a possibilidade de que tal comportamento e topologia crítica sejam características intrínsecas do mecanismo da linguagem humana.

Vale destacar que

- foram analisadas redes semânticas de discursos orais;
- os pesos sobre cada aresta da rede de palavras correspondem aos valores oriundos da equação (3.9) expresso na secção anterior;
- para *FF*= 0, as redes de palavras são constituídas por todas as arestas e vértices;
- o parâmetro de ordem para detecção da rede crítica adotado foi o diâmetro.

Para a presente pesquisa, foi proposto um método semelhante para construção das redes de palavras oriundas de textos literários escritos. No capítulo seguinte, explicitam-se todas as etapas que constituíram este trabalho de pesquisa.

4. O MÉTODO

Esta secção dedica particular atenção às etapas que constituíram esse trabalho de pesquisa. Nele encontram-se informações sobre a base de dados considerada, uma sucinta apresentação dos programas utilizados no tratamento automático e a definição atribuída à distância entre dois elementos dessa amostra.

4.1 A AMOSTRA

> "De resto, é dele [diálogo social] que o enunciado se origina: ele é como a sua continuação, sua réplica, ele não aborda o objeto chegando de não se sabe onde". (BAKHTIN, 1993 apud CUNHA, 2008).

A base de dados desta pesquisa constitui-se de 50 textos literários selecionados de sítios de armazenamento em formato eletrônico e de domínio público. Em geral, foram coletados romances, novelas ou contos por manterem semelhanças quanto à forma da sua narrativa (comunicação dialógica). De acordo com Matta (2004), a linguagem da narrativa inclui formas temporais, causais, intencionais e condicionais. Tais formas ajudam a construir o enredo e, à medida que isto ocorre, relações antes obscuras, ou só parcialmente entendidas, começam a tornar-se inteligíveis.

Segundo Teles (2002), essas três manifestações literárias (o conto, a novela e o romance) enquadram-se na mesma família ficcional e a distinção entre elas "fica quase sempre a critério e preferências do escritor". Alguns estudiosos ainda utilizam como critério de diferenciação a extensão de sua prosa. Porém, qualitativamente, elas apresentam diferenças significativas quanto à estrutura da sua narrativa. Para Burianová (1999), o tratamento do tempo e a descrição das cenas, personagens e suas ações acabam expressando essas diferenças.

> Se reparamos, em termos da problemática temporal, nos principais tipos da prosa de ficção — romance, novela e conto, já pelo simples olhar para a extensão destes gêneros podemos constatar que é o primeiro deles que oferece uma maior flexibilidade no tratamento do tempo. No romance o tempo corresponde à sua heterogeneidade formal, manifestada na capacidade de abranger vários gêneros e procedimentos narrativos. [...] em sentido amplo podemos afirmar que a novela, por causa da sua extensão situada entre o romance e o conto, se distingue do conto por uma maior complexidade formal em termos de digressões e episódios secundários, assim como por um maior espaço dado ao desenvolvimento das personagens.

Diante dessas particularidades, este tipo de narrativa foi escolhido, especificamente, por apresentar um discurso dialógico, pois se compreende que o texto "[...] não é simplesmente uma constelação de palavras, mas sim uma voz narrativa descrevendo eventos reais" (CÔRREA, 2006).

Segundo Cunha (2008), é por meio da linguagem dialógica que se dá a existência do homem. Essa concepção é proposta por Bakhtin que vê na Metalingüística uma disciplina que visa a estudar os "aspectos da vida do discurso" (sic).

Dessa forma, pode-se entender o texto, por exemplo, do romance como uma representação artística da interação humana, interação esta que se expressa através das vozes não só do narrador e das personagens, mas também, do próprio autor (CUNHA, 2008).

> Na teoria bakhtiniana do romance, há dois planos na narração: o do narrador e o do autor, que realiza sua intenção de modo refratado na narração e através dela. O mesmo ocorre com as falas das personagens: elas também podem refratar as intenções do autor, podendo ser a segunda linguagem do autor.

Definido o gênero literário, foram escolhidos quatro idiomas onde três deles têm origem latina (espanhol, francês e português) e um, germânica (inglês). Tais idiomas foram selecionados respeitando dois critérios: disponibilidade (quantidade de textos) e a operacionalidade limitada pela versão do pacote do UNITEX[18] utilizado. Esses textos estão distribuídos conforme Tabela 2.

[18] Unitex é um *software* livre, fruto dos trabalhos desenvolvidos, inicialmente, pelo lingüista Maurice Gross no Laboratoire d'Automatique Documentaire et Linguistique (LADL) e que conta, atualmente, com o apoio da REDE RELEX (um consórcio de Laboratórios Lingüísticos com estudos referentes a diversos idiomas). Seus aplicativos e ferramentas serão discutidos na secção 4.2 desta dissertação.

IDIOMA	QUANTIDADE
Espanhol	8
Francês	13
Inglês	14
Português (Brasil)	15

Tabela 2. Distribuição dos textos literários selecionados quanto à quantidade e idioma
Fonte: Elaborado pela autora, 2009

Vale ressaltar que quatro desses textos estão escritos em três idiomas (francês, inglês e português), são eles:

- *Madame Bovary* de Gustave Flaubert
- *Cinq Semaines en Ballon*, *Le Tour du Monde en Quatre-vingts Jours* e *Voyage au Centre de la Terre* de Jules Verne.

Portanto, o objeto de análise dessa pesquisa é o conjunto de textos literários formado por 42 textos escritos na língua natural de cada escritor e 2 versões relacionadas a cada um dos 4 textos listados acima. Tal conjunto foi submetido, inicialmente, a um pré-tratamento manual, onde cabeçalhos, notas de licença, índices, dentre outros itens, foram retirados com o fim de manter apenas o corpo do texto. Além dessas modificações, foi avaliada a influência das palavras separadas por apóstrofos, como por exemplo: *don't*, that's, *i'd* (para o idioma inglês), d'alma, d'água (para o português) e *l'homme*, *s'est*, *d'un* (para o idioma francês). Tais palavras e contrações estão presentes com certa freqüência tanto na língua inglesa quanto na língua francesa, isto é, representariam *hubs* das redes de palavras dos textos literários nesses idiomas. Com o intuito de eliminar ou diminuir a existência desses casos, realizou-se um segundo pré-tratamento manual em que tais palavras foram modificadas uma a uma, sempre que possível. Somente foram mantidas aquelas palavras que ocorreram poucas vezes no texto (baixa freqüência), visto que elas não apresentariam associações significativas.

Para facilitar a compreensão dessa fase, segue um trecho do texto *The Chimes* de Charles Dickens:

"[...]***That's*** *the fact. He* ***didn't*** *seem to wait so long for a sixpence in the wind, as at other times*;[...]"

Com o pré-tratamento, o trecho acima se torna:

"[...]***That is*** *the fact. He* ***did not*** *seem to wait so long for a sixpence in the wind, as at other times;*[...]"

Após essa etapa, os textos foram salvos (extensão .txt) segundo uma regra de nomenclatura:

- caracteres do idioma (letra maiúscula)_caracteres do autor (letra maiúscula)_qualquer palavra do título do texto (letra minúscula)

Os Quadro 1e Quadro 2 mostram, em ordem alfabética, a simbologia usada para definir idioma e autor.

CARACTERES	IDIOMA
ES	Espanhol
FR	Francês
IN	Inglês
PT	Português (Brasil)

Quadro 1. Significado da primeira posição do nome do arquivo: idioma
Fonte: Elaborado pela autora, 2009

CARACTERES	AUTOR(A)	CARACTERES	AUTOR(A)
AA	Aluísio de Azevedo	JA	Jane Austen
AD	Alexandre Dumas	JA	José de Alencar
BG	Benito Pérez Galdós	JV	Juan Valera
CD	Charles Dickens	JV	Jules Verne
EA	Edmond About	MA	Machado de Assis
GA	Guillaume Apollinaire	PF	Paul Féval
GF	Gustave Flaubert	VI	Vicente Blasco Ibánes
JA	Jean Aicard		

Quadro 2. Significado da segunda posição do nome do arquivo: autor
Fonte: Elaborado pela autora, 2009

Por exemplo,

PT_JA_gazela significa que o arquivo encontra-se em Português, foi escrito por José de Alencar e refere-se ao romance *A Pata da Gazela*.

No APÊNDICE A encontra-se uma tabela com a composição de todos os textos utilizados e suas classificações.

Em seguida, os textos passaram pelo processo de pré-tratamento automático utilizando uma série de programas oriundos do pacote UNITEX e dos trabalhos de Caldeira (2005) e Teixeira (2007). Como produto deste processo, pode-se construir e visualizar a rede semântica gerada pelas associações de palavras que constituem as frases do texto literário a partir do

software PAJEK[19]. A análise e apresentação desses resultados foram realizadas utilizando o software Origin[20].

Na tentativa de se criar classes para avaliar os resultados obtidos da amostra analisada, agrupou-se os textos segundo três atributos: idioma (I), conteúdo (C) e autor (A). Desses três atributos, foi possível promover oito possíveis combinações. Apesar de ser apresentado todas as combinações e suas respectivas interpretações, é importante ressaltar que três delas não fazem sentido em serem analisadas, são: (a), (c) e (g). Além disso, a combinação (f) apresenta duas variáveis diferentes, de forma que não seria possível avaliar se a estrutura topológica extraída da rede seria resultado da influência do próprio autor ou do tradutor. Devido a isso, a combinação (f) também foi eliminada da presente análise.

Abaixo, seguem as descrições:

(a) mesmo idioma, mesmo conteúdo e mesmo autor → um único texto representado por ele mesmo (o texto em si);

Exemplo: *Marianela* escrito em espanhol por Benito Pérez Galdós

(b) mesmo idioma, conteúdo diferente e mesmo autor → textos diferentes escritos num mesmo idioma e por um mesmo autor (avaliação do conjunto composto pelos livros originais de um autor);

Exemplo: *Hard Times*, *Mugby Junction* e *The Chimes* escritos em inglês por Charles Dickens

(c) mesmo idioma, mesmo conteúdo e autor diferente → textos iguais escritos num mesmo idioma e por autores diferentes (como se existissem dois livros iguais escritos por autores diferentes);

(d) mesmo idioma, conteúdo diferente e autor diferente → textos diferentes escritos por autores diferentes e num mesmo idioma (avaliação dos textos literários escritos num único idioma);

Exemplo: *Madame Bovary* por Gustave Flaubert, *Notre-Dame D'Amour* por Jean Aicard e *Le Loup Blanc* por Paul Féval, todos escritos em francês

(e) idioma diferente, mesmo conteúdo e mesmo autor → textos iguais escritos em

[19] O programa PAJEK é um programa de código aberto para o sistema operacional Windows que foi desenvolvido para permitir a criação, manipulação e visualização de grafos de qualquer tamanho. Ele está disponível em http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[20] O Origin é uma ferramenta produzida pela OriginLab Corporation, comumente utilizada em Física, para análise de dados e gráficos técnicos.

idiomas diferentes (versões/traduções de um mesmo texto);

Exemplo: *Cinq Semaines en Ballon*, *Le Tour du Monde en Quatre-vingts Jours* e *Voyage au Centre de la Terre* de Jules Verne escritos, originalmente, em francês e suas versões em inglês e português

(f) idioma diferente, conteúdo diferente e mesmo autor → textos diferentes escritos em idiomas diferentes por um mesmo autor (avaliação do conjunto composto pelos livros de um autor escritos em diversos idiomas);

Exemplo: *Cinq Semaines en Ballon* escrito em francês (original), *Le Tour du Monde en Quatre-vingts Jours* em inglês (Around the World in Eighty Days) e *Voyage au Centre de la Terre* em português (Viagem ao Centro da Terra), todos de Jules Verne

(g) idioma diferente, mesmo conteúdo e autor diferente → nesse caso, tem-se uma situação semelhante a aquela representada pela letra (c) (como se existissem dois livros iguais escritos por autores diferentes);

(h) idioma diferente, conteúdo diferente e autor diferente → os mais variados textos para os quais todos os critérios acima devem ser diferentes.

Exemplo: *La Barraca* em espanhol por Vicente Blasco Ibanés, *Bric-a-Brac* em francês de Alexandre Dumas, *Persuasion* em inglês de Jane Austen e *Diva*, escrito em português por José de Alencar.

Apesar de ser possível a análise da combinação (h), ter-se-ia que considerar um número maior do que quatro idiomas para que os resultados fossem estatisticamente relevantes[21]. As demais combinações (b), (d) e (e) representam, respectivamente, o autor, idioma e conteúdo e compõem as classes de agrupamentos dos textos.

Dessa forma, cada classe será composta por um número fixo de textos que satisfaçam às especificidades descritas anteriormente. Dentro de cada classe, foi avaliada a distância euclidiana entre os vetores definido pelos índices de rede. Considerando cada texto como um ponto no espaço, e supondo que, a partir desses índices, fosse possível extrair características relacionadas à linguagem humana quanto ao idioma, conteúdo e autor, pode-se enunciar duas proposições:

1) Dentro de cada classe, existem grupos de textos em que a distância euclidiana

[21] Com isso, não se quer dizer que a amostra que foi analisada neste trabalho seja estatisticamente significantes, visto que este corresponde a um trabalho de análise do método. A mesma justificativa emprega-se à escolha e

entre eles é menor do que a distância entre textos pertencentes à mesma classe, mas que não pertençam ao mesmo grupo.

Por exemplo: Se tomar a classe AUTOR e calcular-se a distância entre os textos de Jules Verne, esta distância deverá ser menor do que a distância entre um texto de Jules Verne e qualquer outro autor.

Assim, cada classe foi composta por grupos internos no qual cada um deles representa um autor, um idioma ou um conteúdo;

2) Caso não haja a formação desses grupos para ao menos uma dessas classes, então se está identificando algo que se aproxima da proposta de S. Pinker e que ficou conhecida como mentalês[22]. Isto significa que, como a linguagem é própria da espécie humana, não há diferenças significantes que caracterizem estatisticamente a rede de palavras oriundas de um texto literário.

Na secção 4.4, esclarecem-se os processos envolvidos na análise desta parte da pesquisa.

4.2 TRATAMENTO DOS DADOS

Após o tratamento manual realizado e descrito na secção 4.1, os textos selecionados foram submetidos a um tratamento automático constituído, inicialmente, por um conjunto de programas oriundos do UNITEX e, posteriormente, pelos programas desenvolvidos por Caldeira (2005) e modificados por Teixeira (2007).

O UNITEX (UNITEX, 2002) é um software livre de tratamento de textos em língua natural constituído de certos recursos lingüísticos como: dicionários eletrônicos, gramáticas e tábuas léxico-gramaticais. Tais recursos permitem tratar diversos sistemas de escrita avaliando-os tanto em níveis morfológicos como sintáticos.

Por ser um conjunto de programas livremente distribuído, o UNITEX vem sofrendo

---

quantidade dos índices de rede.

[22] O mentalês, tese fundamentada na teoria de seleção natural de Charles Darwin e na gramática universal de Noam Chomsky exposta por Steven, é uma espécie de código mental inato ao ser humano e, portanto, “uma peça de constituição biológica do cérebro” (PEREIRA, 2002). Assim, para Pinker, a linguagem é uma habilidade complexa e especializada que se desenvolve espontaneamente e que é qualitativamente a mesma em todo indivíduo

modificações desde a sua versão inicial (versão 1.0). Esta versão incluía fontes para os idiomas: francês, grego, inglês, português e tailandês.

As versões seguintes (1.1, 1.2 e 2.0), além de incluírem outros idiomas, sofreram mudanças e correções de *bugs*. A versão 1.2 (2006) e a mais atual (versão 2.0) incluem em seus pacotes dicionários capazes de tratar textos escritos em alemão, coreano, espanhol, finlandês, francês, grego, inglês, italiano, norueguês, polonês, português, russo, sérvio e tailandês.

No que tange a esta pesquisa, os dicionários eletrônicos assumem um papel de fundamental importância, visto que um dicionário pouco "estruturado" causaria problemas na identificação das palavras que compõem os textos. Devido a isso, utilizou-se a versão 1.2 que está parcialmente[23] preparada para a análise de textos.

Em geral, tais ferramentas do UNITEX são elaboradas por equipes de lingüistas para as mais variadas línguas e estão representados com o formalismo DELA (Dicionários Eletrônicos LADL). Esse formalismo possibilita descrever as entradas lexicais (as palavras) simples e compostas de uma língua associando opcionalmente informações gramaticais, semânticas ou flexionais.

O Quadro 3 apresenta, resumidamente, alguns códigos gramaticais usados nos dicionários fornecidos pelo UNITEX.

CÓDIGO	SIGNIFICADO	EXEMPLOS
A	Adjetivo	fabuloso
ADV	Advérbio	ontem, de repente
CONJC	Conjunção de Coordenação	mas
CONJS	Conjunção de Subordinação	embora, a menos que
DET	Determinante	uma, seus, vinte
INTJ	Interjeição	tchau
N	Substantivo	mesa, bolsa de valores
PREP	Preposição	sem, à margem de
PRO	Pronome	ela, a gente
V	Verbo	cantar, ver

Quadro 3. Códigos gramaticais usuais do UNITEX
Fonte: Manual Unitex (2002)

[23] O motivo da palavra "parcialmente" neste parágrafo é devido a problemas verificados no decorrer deste trabalho nos dicionários do UNITEX. Tais problemas somaram-se a alguns outros quanto à escolha dos textos que formaram a base de dados dessa pesquisa.

É importante salientar que, apesar de existir uma codificação comum para a maioria das línguas, os dicionários contêm especificidades próprias de cada língua. Assim, em caso de dúvida quanto à codificação, sugere-se contato com o próprio autor do dicionário ou verificação, posterior à execução, dos arquivos gerados por um aplicativo do UNITEX chamado *Dico*[24].

Tendo em vista essas particularidades associadas ao UNITEX, foram geradas quatro pastas contendo todos os arquivos correspondentes a cada idioma utilizado nesta pesquisa. Ou seja, em cada pasta tem-se um conjunto de dicionários, gramáticas e tábuas léxico-gramaticais próprio para cada idioma e aplicativos comuns aos quatro idiomas. A Figura 20 representa uma destas pastas composta pelo mínimo de elementos necessários para o tratamento dos textos.

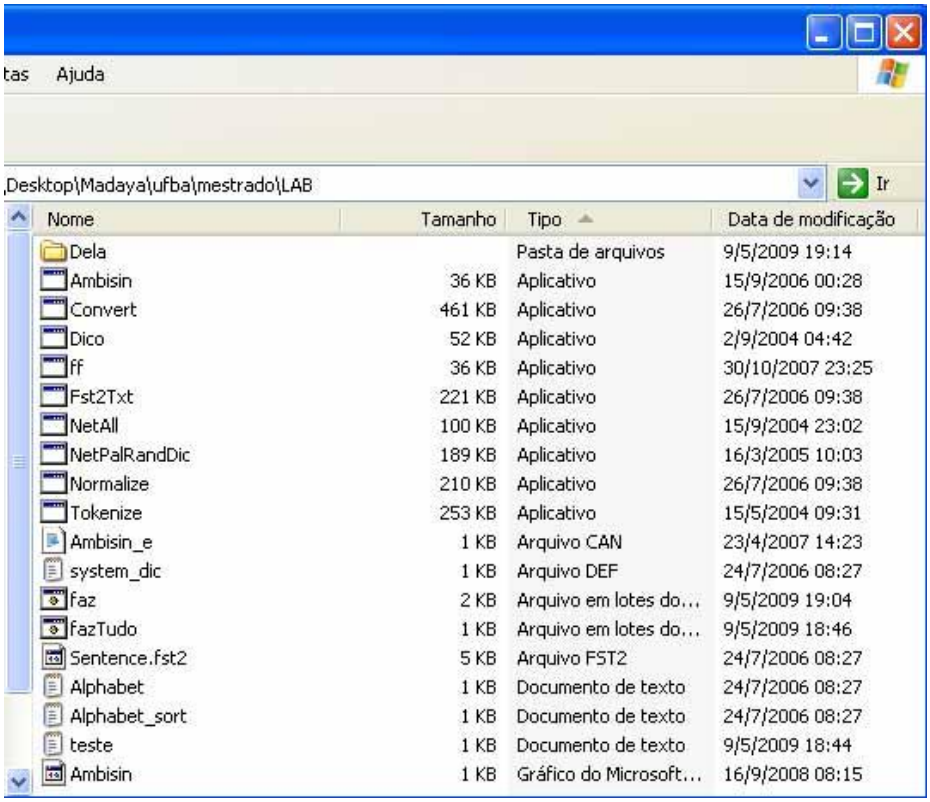


Figura 20. Representação de uma pasta 'LAB' que contém o número mínimo de elementos necessários para o tratamento de um texto nomeado por 'teste'.
Fonte: Elaborado pela autora, 2009

É importante evidenciar que, destes arquivos e aplicativos, nove foram retirados diretamente do UNITEX. São eles: *Alphabet* e *Alphabet_short*, *Convert*, *Dela*, *Fst2txt*,

[24] Informações sobre este aplicativo serão, resumidamente, apresentadas ainda no decorrer deste capítulo.

Normalize, *Sentence*, *system_dic* e *Tokenize*.

Os programas de lotes vêm sofrendo modificações desde o trabalho de Caldeira (2005). Nesta pesquisa, fazem parte desses programas os arquivos *faz* e *fazTudo*. O arquivo f*azTudo.bat* é o responsável por chamar o arquivo de lote *faz.bat* para todos os arquivos de textos descritos na variável *va* (ressaltada na Figura 21).

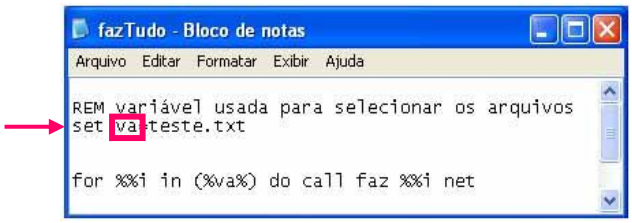


Figura 21. Arquivo de lote *fazTudo.bat*
Fonte: Adaptado de Teixeira (2007)

A Figura 22 representa o diagrama da primeira versão do programa de lote *faz.bat*.

Criação de pastas e subpastas para organizar os diversos arquivos gerados	*mkdir Cres_%1* cd Cres_%1 *mkdir %1_snt*
Conversão do texto ASCII para Unicode, [e transferência do arquivo gerado para a subpasta criada]	..\asc2uni PORTUGUESE ..\%1 move ..\%1.uni .\
Normalização de Separadores [elimina tabulação, retorno de linha (enter) e espaço excedentes]	..\normalize %1.uni
Segmentação em sentenças, [usando arquivo *Alphabet* correspondente ao idioma do texto]	..\Fst2Txt.exe %1.snt ..\Sentence.fst2 ..\Alphabet.txt -merge
Segmentação em unidades lexicais	..\Tokenize.exe %1.snt ..\Alphabet.txt
Normalização das formas não-ambíguas [e aplicação dos dicionários]	..\dico %1.snt ..\Alphabet.txt ..\Dela\Delaf_pb.bin
Conversão do texto Unicode para ASCII	..\uni2asc PORTUGUESE %1_snt\dlf
Eliminação das ambigüidades, [escolha das palavras que comporão a rede e eliminação das palavras gramaticais]	..\Ambisin %1_snt\dlf.ascii %1_snt\dlf.txt 2

Figura 22. Diagrama do pré-tratamento dos textos e linhas do código do arquivo BAT usado para chamar os programas.
Fonte: Caldeira (2005)

Na Figura 23, apresenta-se a versão mais recente deste arquivo BAT[25] considerando as modificações realizadas no trabalho de Teixeira (2007) e aquelas oriundas da versão do UNITEX adotada nessa pesquisa. Note que alguns programas deste lote, originados do pacote UNITEX, são comuns a ambas as versões e mantêm a mesma função.

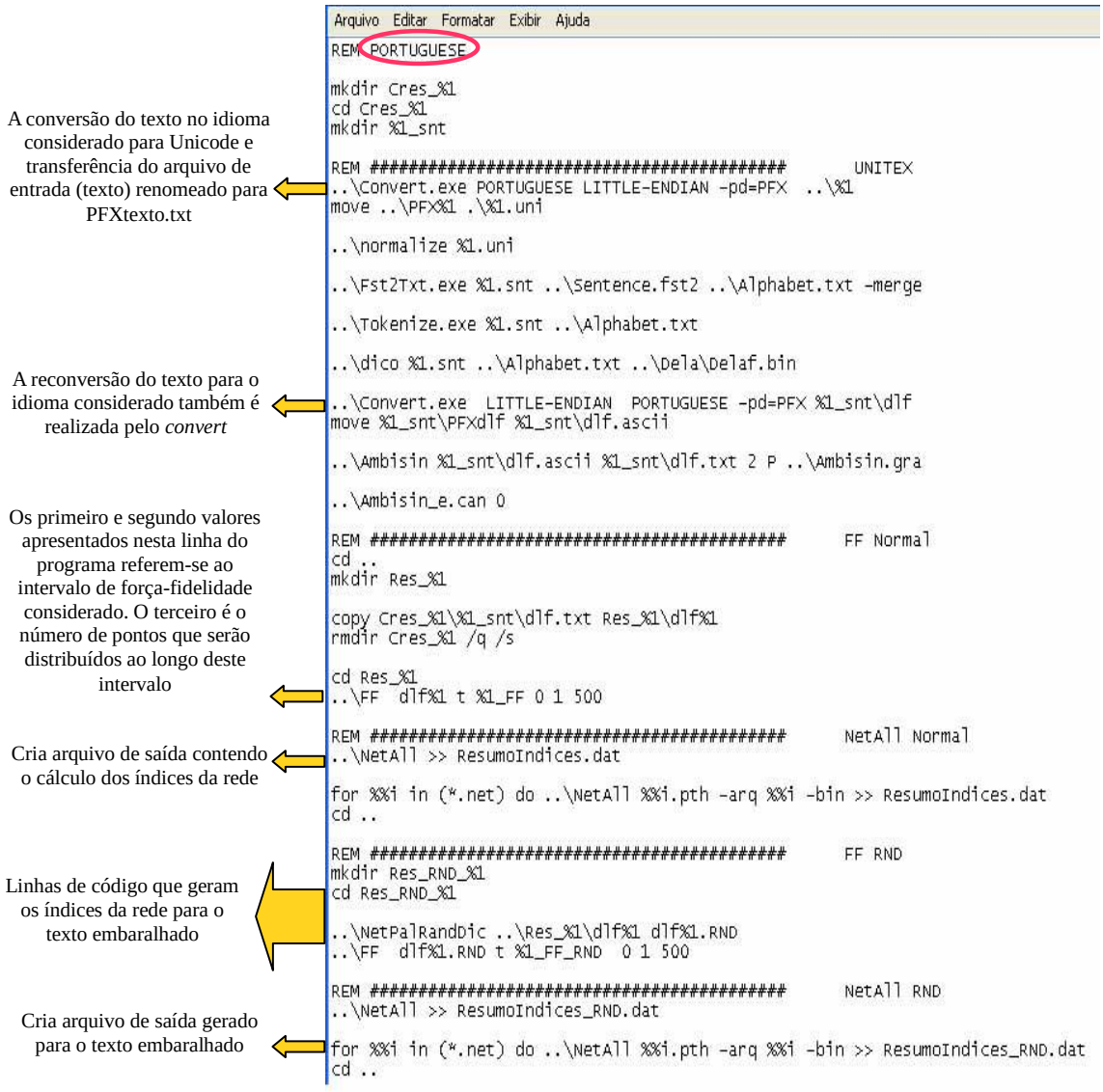


Figura 23. Diagrama do arquivo de lote *faz.bat* usado para chamar os programas para tratamento automático dos textos.
Fonte: Elaborado pela autora, 2009

[25] Este arquivo varia de idioma para idioma devido às suas especificidades, haja vista que as funções de dicionários que são chamadas em suas linhas de código dependem do idioma considerado. Na Figura 23, é considerado o idioma português. Para construir um programa de lotes para a língua inglesa, por exemplo, basta substituir o parâmetro 'PORTUGUESE' por 'ENGLISH'. O UNITEX 1.2 apresenta uma lista de parâmetros usados nesta versão (UNITEX 1.2, 2006)

Para fins de esclarecimento, segue-se uma brevíssima síntese sobre algumas funções atribuídas a cada linha do arquivo BAT correspondente ao pacote UNITEX.

(a) Durante o tratamento automático inicial, são criadas pastas (comando *mkdir*) onde são salvos os arquivos para posterior manipulação. Essas pastas são nomeadas por *RES_nome do arquivo* (para o texto original) e *RES_RND_nome do arquivo* (para o texto embaralhado).

(b) Com o programa *Convert*, textos escritos no formato padrão (ASCII) são convertidos de certo idioma em linguagem Unicode (*Little-Endian*) e vice-versa. Ele substitui os programas *Asc2Uni* e *Uni2Asc* das versões anteriores.

(c) *Normalize* é um programa que normaliza o texto buscando tanto a identificação das palavras que o constituem quanto a eliminação dos separadores de textos (espaços, tabulações e *enter*) e a delimitação das frases. Essa delimitação das frases é uma parte muito importante no processo de tratamento dos textos, pois é a frase que representa a menor unidade de significação, a idéia. Este processo de delimitar frases ocorre a partir da identificação de uma seqüência de símbolos separados no texto. Essas frases são delimitadas por {S}. Como resultado dessa normalização, tem-se um arquivo com extensão .snt.

A Figura 24 mostra um exemplo de texto (*Quadrilha*, de Carlos Drummond de Andrade) antes e depois de passar pelo tratamento desse programa e a criação do arquivo .snt.

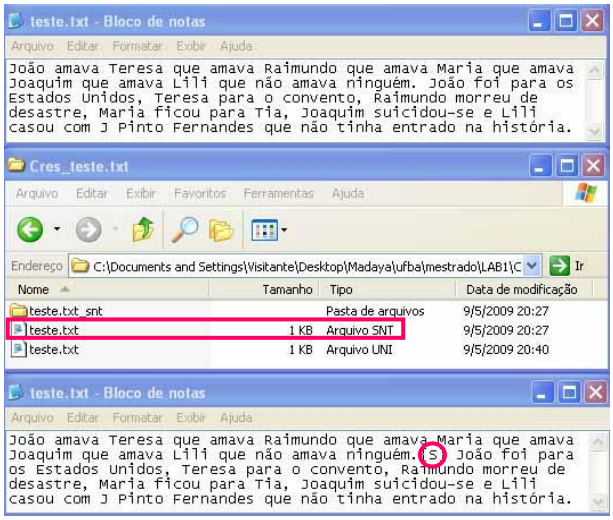


Figura 24. Ilustração que mostra o produto do tratamento de um texto 'teste' obtido da execução do programa normalize.
Fonte: Elaborado pela autora, 2009

(d) O programa *Fst2Txt* aplica um transdutor (fst2) ao texto no estágio anterior ao corte do texto em unidades lexicais. Alguns dos parâmetros deste programa são: *fst2*, *alph*, *mode*. O parâmetro *alph* aplica os arquivos relacionados ao alfabeto e *mode* as duas modalidades possíveis do transdutor (*-merge* e *-replace*). *Sentence*.fst2 é a gramática aplicada ao texto pelo programa *Fst2Txt* na modalidade *merge*.

(e) A segmentação do texto em unidades lexicais é realizada a partir do programa *Tokenize*. Ele cria vários arquivos no diretório para armazenar informações sobre o texto (*Cres_nome do arquivo*). São eles: *tokens.txt*, *text.cod*, *tok_by_freq.txt*, *stats.n*, *tok_by_alph.txt*.

(f) Apesar do *Dico* ser um programa do UNITEX, ele foi modificado durante o trabalho de Caldeira (2005) para atender os objetivos de sua pesquisa. Esta alteração visava manter, no arquivo dlf.ascii, palavras que não estavam armazenadas no dicionário indicando-as pela nomeação 'NOTFOUND' conforme a Figura 25. Dessa forma, além de ser feita a aplicação do recurso do dicionário, isto é, as unidades gramaticais são classificadas e aquelas consideradas como verbos são reescritas na sua forma canônica, as palavras desconhecidas pelo programa não serão eliminadas do texto.

Da execução deste programa, podem ser gerados, no diretório do texto analisado, três arquivos mostrados no Quadro 4.

ARQUIVO	DESCRIÇÃO
dlf	Para palavras simples
dfl	Para palavras compostas
err	Para palavras desconhecidas

Quadro 4. Arquivos produzidos pelo programa *Dico*
Fonte: Adaptado de Teixeira (2007)

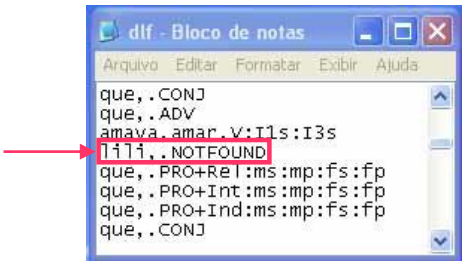


Figura 25. Exemplo de um arquivo dlf.ascii de um texto.
Fonte: Elaborado pela autora, 2009

Como já foi dito anteriormente, a aplicação do dicionário usa o formato DELA (Dicionários Eletrônicos LADL) que permite a descrição das entradas lexicais simples e compostas de um idioma relacionando-as com informações gramaticais, semânticas e flexionais. Para classificar tais informações, foram utilizados dois tipos distintos de dicionários eletrônicos: o DELAF (dicionário de informações flexionadas simples ou compostas) e o DELAS (dicionário de formas canônicas).

(g) O Ambisin é um programa desenvolvido no trabalho de Caldeira (2005) com o intuito de eliminar palavras gramaticais, minimizar efeitos de ambigüidades e separar as formas flexionadas ou canônicas das palavras do restante dos itens de classificação gramatical gerada pelo UNITEX (CALDEIRA, 2005). Na pesquisa acima citada, foram propostos quatro parâmetros a serem utilizados por este executável como mostra o Quadro 5.

PARÂMETRO	DESCRIÇÃO
0	Mantém flexões e não exclui palavras gramaticais
1	Reduz as palavras em sua forma canônica e não exclui palavras gramaticais
2	Reduz as palavras em sua forma canônica e exclui palavras gramaticais
3	Mantém flexões e exclui palavras gramaticais

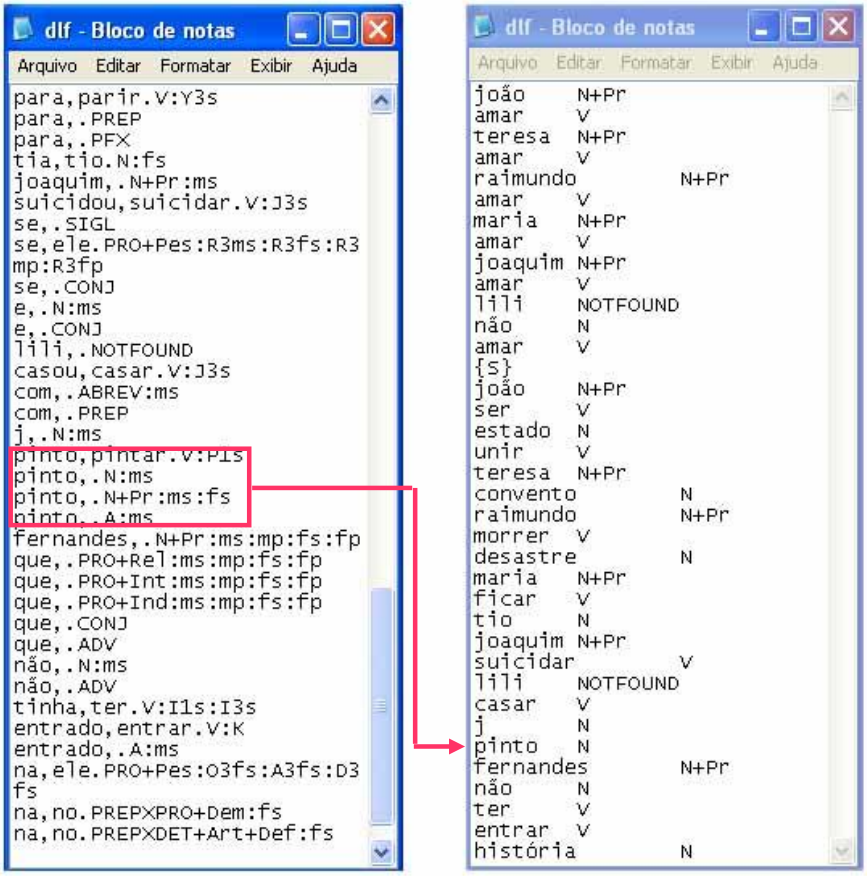
Quadro 5. Parâmetros que podem ser usados no programa *Ambisin*
Fonte: Adaptado de Teixeira (2007)

Assim como nos trabalhos de Caldeira (2005) e Teixeira (2007), o parâmetro escolhido foi o 2[26]. Isto significa que após realizada a classificação gramatical e a exclusão de classes gramaticais (mediada pelo arquivo de filtragem *Ambisin.gra*), é ordenada a precedência de palavras ambíguas[27] considerando a seqüência: substantivo, verbo, adjetivo dentre outros.

Outro arquivo de filtragem é o *Ambisin_e.can.* Ele permite que palavras especificadas previamente sejam excluídas. Tais palavras e classes gramaticais são eliminadas pelo programa *Ambisin* por serem consideradas signos que não trazem consigo carga semântica.

Para ilustrar como se dá o processo de ordenação, seja a palavra 'pinto' apresentada na Figura **26**.

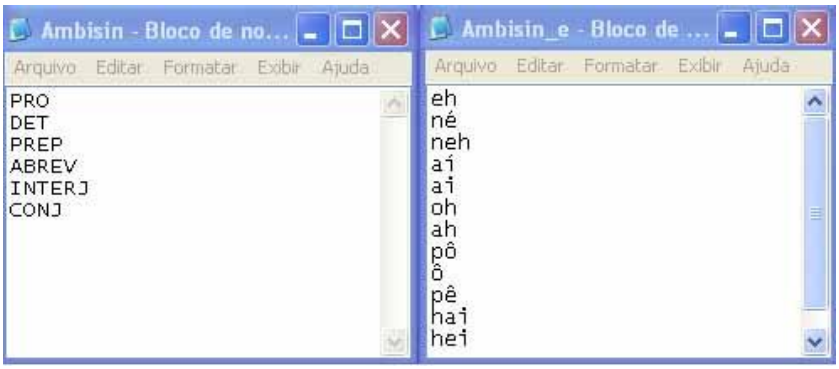
[26] Este tratamento foi escolhido, pois está relacionado com uma estrutura de associações diferente dos outros tratamentos. Para maiores esclarecimentos, ver Caldeira (2005).
[27] Ambigüidades são, em geral, grandes problemas de softwares como o UNITEX. Elas ocorrem quando uma ou mais palavras apresentam várias classificações gramaticais.



(A) (B)

Figura 26.. Ilustração da aplicação da ordem de precedência na classificação gramatical de palavras realizada pelo *Ambisin* onde, no arquivo dlf.ascii (A), 4 classificações gramaticais são listadas sendo que uma delas é o substantivo (N). Então, pela ordem de precedência, essa é a classe gramatical escolhida e apresentada no arquivo dlf.txt (B).
Fonte: Elaborado pela autora, 2009

A Figura 27 ilustra um exemplo aplicado ao arquivo *Ambisin*.gra e *Ambisin_e.can*.



(A) (B)

Figura 27. Exemplo aplicado ao *Ambisin.gra* (A) e *Ambisin_e.can* (B). Adaptação: Teixeira (2007)
Fonte: Elaborado pela autora, 2009

4.3 CONSTRUÇÃO DA REDE DE PALAVRAS

Após o tratamento automático do texto, tem-se estabelecida a associação entre cada nó da rede e uma palavra que compõe o texto em análise e sobre cada aresta um valor de FF_N correspondente. Ou seja, é possível determinar a Força-Fidelidade para cada pare de palavras e, posteriormente, os índices característicos da rede complexa correspondentes a cada valor de Força-Fidelidade.

Tais índices e freqüências são calculados por programas de código livre, chamados, respectivamente, por *NetAll* e *FF*, que foram desenvolvidos em colaboração durante o trabalho de Caldeira (2005), Teixeira (2007) e nesta pesquisa.

Os arquivos de extensão .freq e .net são resultados da execução do aplicativo *FF*. Este executável calcula a Força-Fidelidade dos pares de palavras do texto analisado, conforme as equações mostradas na secção 3.3 do capítulo anterior, e gera a rede de palavras a partir destes valores.

Ou seja, após o tratamento automático realizado pelo UNITEX, o arquivo de lote *faz.bat* executa este programa que calcula vários valores de Força-Fidelidade considerando o intervalo de Força-Fidelidade e o número de pontos desejado[28].

A Figura 28 corresponde à parte do arquivo de saída .freq para o texto *Quadrilha.*

Total de Frases: 2								
Vocabulário: 24								
Voc/Sen: 12								
PAR	#SENT1	#SENT2	FREQPAR	FORCA	FORCAN	FID.	FIDN	FF
joão-amar	2	1	1	0.5	0	0.5	0	0
joão-teresa	2	2	2	1	1	1	1	1
amar-teresa	1	2	1	0.5	0	0.5	0	0
joão-raimundo	2	2	2	1	1	1	1	1
amar-raimundo	1	2	1	0.5	0	0.5	0	0
teresa-raimundo	2	2	2	1	1	1	1	1
joão-maria	2	2	2	1	1	1	1	1
amar-maria	1	2	1	0.5	0	0.5	0	0
teresa-maria	2	2	2	1	1	1	1	1
amar-joaquim	1	2	1	0.5	0	0.5	0	0
maria-joaquim	2	2	2	1	1	1	1	1

Figura 28. Ilustração de parte do arquivo .freq para o texto *Quadrilha* (original)
Fonte: Elaborado pela autora, 2009

[28] Estes dados são atribuídos pelo usuário.

Este programa utiliza como entrada de dados o arquivo dlf_.txt, para o texto original, e dlf_.RND, para o texto embaralhado. Assim, um arquivo semelhante ao exposto acima também é gerado para o texto embaralhado.

Os arquivos com extensão .net são utilizados pelo PAJEK para a visualização das redes de palavras que compõem o texto analisado (Figura 29). Dessa forma, para cada valor de Força-Fidelidade existe um arquivo .net correspondente em que é possível visualizar a rede cujos pesos das arestas são valores maiores ou iguais ao valor de FF_N.

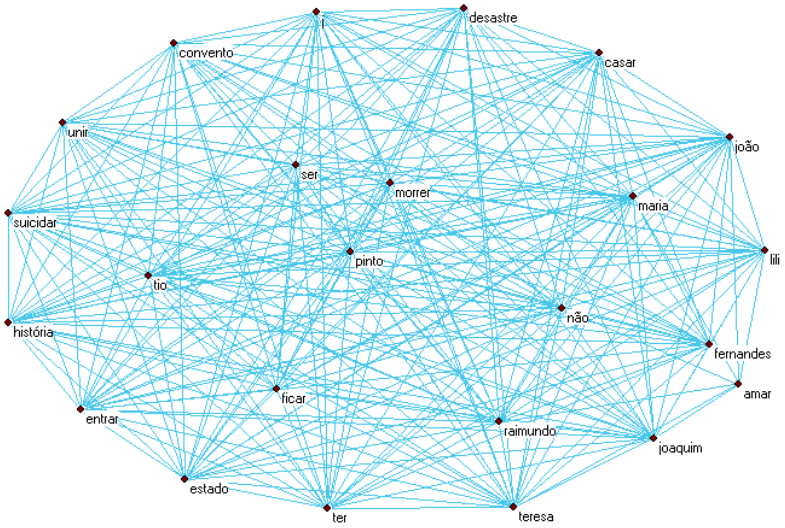


Figura 29. Ilustração da rede de palavras do texto *Quadrilha* para o valor de $FF_N = 0$.
Fonte: Elaborado pela autora, 2009

A determinação da Força-Fidelidade é fundamental para o cálculo dos índices característicos da rede. Portanto, a execução do *FF* deve ser anterior à do *NetAll*.

A sintaxe para execução do programa *NetAll* é mostrada na Figura 30.

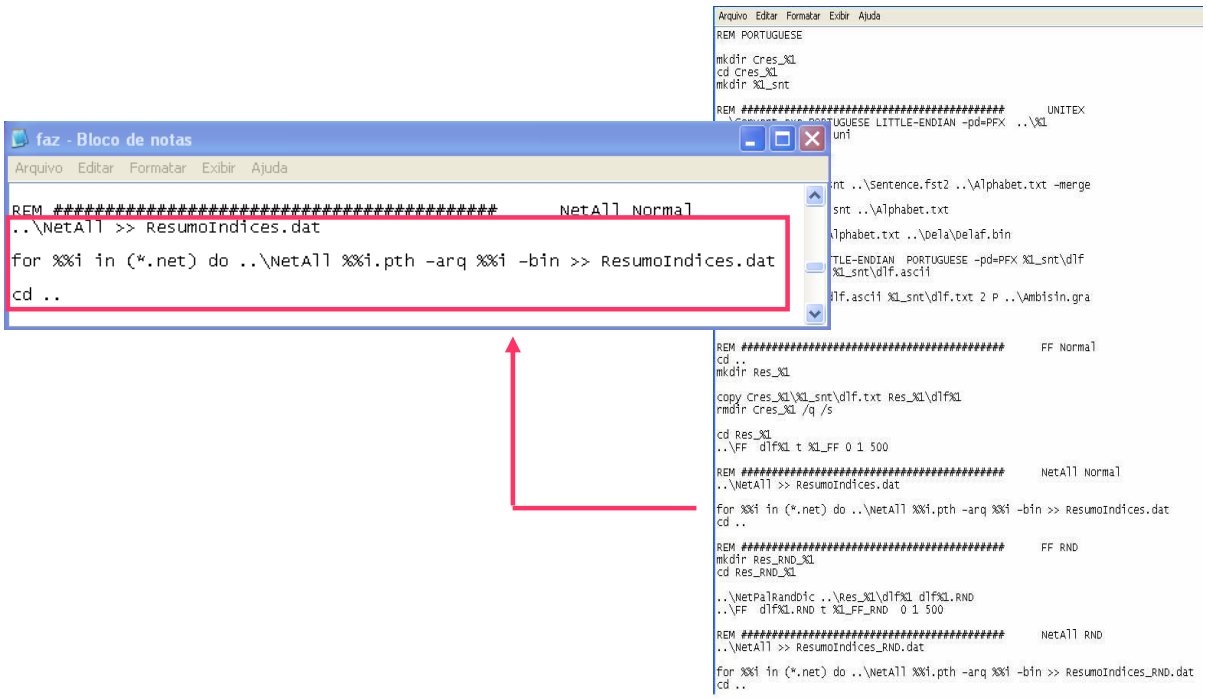


Figura 30. Zoom do arquivo de lotes *faz.bat* destacando a sintaxe para execução do programa *NetAll*. (ver Figura 23)
Fonte: Elaborado pela autora, 2009

Os parâmetros de rede apresentados na linha de código em destaque na Figura 30 fazem parte de um grupo de parâmetros relacionados ao tipo de rede. O Quadro 6 elenca os tipos de redes e os possíveis parâmetros correspondentes.

TIPOS DE REDE	PARÂMETROS
-arq: Rede em um arquivo do tipo pajek (.net)	Parâmetro 1: Nome do arquivo de entrada Parâmetro 2: Tipo de cálculo
-rnd: Rede do tipo aleatória a partir de uma preexistente	Parâmetro 1: Nome do arquivo de entrada Parâmetro 2: Tipo de cálculo
-nrd: Gera uma nova rede aleatória	Parâmetro 1: Número de nós na rede Parâmetro 2: Tipo de cálculo Parâmetro 3: Probabilidade, entre 0 e 1, associada à rede aleatória
-tri: Gera uma rede tridiagonal	Parâmetro 1: Número de nós na rede Parâmetro 2: Tipo de cálculo
-srf: Gera uma rede do tipo Livre de Escala (*Scale Free*)	Parâmetro 1: Número de nós na rede Parâmetro 2: Tipo de Cálculo

Quadro 6. Parâmetros do programa *NetAll*
Fonte: Adaptado de Teixeira (2007)

Os tipos de cálculos mencionados na tabela acima são:

- -bin: caminho mínimo mediante produto matricial binário
- -amo: caminho mínimo médio por amostragem sendo o parâmetro 3 o percentual
- -img: gera uma imagem com as matrizes de ordem superior

Para gerar os resultados relacionados ao texto embaralhado são utilizados os mesmos parâmetros acima listados, porém o texto já deve ter sido submetido ao processo de embaralhamento mediado pelo programa *NetPalRandDic*. Neste processo, o número de frases e o de palavras que compõem cada frase do texto são mantidos em relação ao texto original, porém o vocabulário que constitui cada uma dessas frases é escolhido aleatoriamente dentro do conjunto de palavras deste texto (Figura 31). Isto resulta uma alteração da frequência das palavras do texto. Ou seja, este tipo de embaralhamento foi usado por ser o único capaz de quebrar a estrutura de uma rede do tipo livre de escala (CALDEIRA, 2005).

(A) (B)

Figura 31. Ilustração dos arquivos .txt (A) e .RND (B) referentes à primeira frase oriunda, originalmente, do poema *Quadrilha* e após ele ter passado pelo processo de embaralhamento.
Fonte: Elaborado pela autora, 2009

Os arquivos de saída, onde estão armazenados os índices de redes relacionados ao texto original e ao embaralhado, são, respectivamente, ResumoIndices.dat e ResumoIndices_RND.dat.

Além destes arquivos de saída e os mencionados anteriormente, são gerados arquivos com extensão .net.pth. Nestes arquivos estão registradas as quantidades relacionadas ao coeficiente de aglomeração, caminho mínimo médio e grau para cada nó (palavra) da rede referente para cada valor de Força-Fidelidade.

Em geral, os índices calculados neste trabalho são os mesmos estudados por Caldeira (2005) e Teixeira (2007): número de vértice, número de arestas, diâmetro, coeficiente de aglomeração, caminho mínimo médio e grau médio.

Para promover a comparação entre as redes de palavras a partir das classes citadas na secção 4.1 deste capítulo, foi utilizado o conceito de distância euclidiana entre redes no espaço dos índices característicos da Rede Complexa. Este procedimento será explicitado com um pouco mais de detalhe na secção a seguir.

4.4 DETERMINAÇÃO DA DISTÂNCIA EUCLIDIANA ENTRE TEXTOS

Com a finalidade de se verificar a formação de grupos dentro de cada uma das três classes, selecionou-se 36 textos que foram, convenientemente, agrupados de forma que cada uma das classes analisadas (autor, conteúdo e idioma) contenha o mesmo número de textos.

Utilizando os índices de Redes Complexas extraídos dos arquivos de saída do *NetAll*, procedeu-se a comparação das redes dos textos constituintes de cada classe a partir da distância euclidiana entre eles no espaço dos índices.

Essa distância é dada por

$$\delta_{i,j} = \left[\sum_{m=1}^{N} (I_{mi} - I_{mj})^2 \right]^{1/2} \tag{4.1}$$

em que, para o contexto dessa pesquisa,

$\delta_{i,j}$ é a distância entre dois textos distintos i e j

N corresponde ao número total de índices característicos

I_{mi} é a m-ésima coordenada, no espaço dos índices, da rede referente ao texto i

I_{mj} é a m-ésima coordenada, no espaço dos índices, da rede referente ao texto j

É importante salientar que estas coordenadas, bem como a distância $\delta_{i,j}$, sofreram uma

transformação com o fim de que seus valores fossem estendidos de forma a preencher todo o intervalo [0,1].

Dos resultados obtidos, avaliou-se, dentro de cada classe, as diferenças médias entre as distâncias dos textos que possuem a mesma característica que define a classe (intragrupos) e os demais textos pertencentes à classe (intergrupos), a partir de um teste paramétrico para amostras independentes – Teste T[29].

Para esclarecer esse procedimento, considere a Tabela 3 como uma ilustração representando quatro textos que foram agrupados na classe AUTOR. Note que esta tabela contém informações referentes à numeração de cada grupo, texto adotado, Força-Fidelidade normalizada e os índices de rede correspondentes a estes valores.

NUMERAÇÃO DO GRUPO	TEXTO	FFn	D	CAM	CMM	<k>	γ
1	ES_VI_arroz_tartana	3.47×10^{-4}	17	0.15	4.47	4.9	1.69
1	ES_VI_catedral	2.98×10^{-4}	16	0.2	3.94	6.4	1.81
2	FR_AD_bric_a_brac	3.72×10^{-4}	14	0.26	4.07	5.61	1.7
2	FR_AD_femme	5.20×10^{-4}	15	0.23	3.99	6.76	1.76

Tabela 3. Exemplo de uma tabela, considerando apenas 2 autores, contendo as informações necessárias para calcular a distância euclidiana entre textos pertencentes a uma mesma classe
Fonte: Elaborado pela autora, 2009

A numeração do grupo diferencia os textos correspondentes ao autor (1) e (2).

Após calcular-se a distância entre textos distintos, para todas as possíveis combinações, realizou-se um teste T, com significância α=0.05, para avaliar as diferenças médias existentes entre os textos pertencentes ao intragrupo, formado pela combinação do tipo (1)-(1) e (2)-(2), e intergrupo, formado pelas combinações do tipo (1)-(2). Isto é, as combinações entre:

- ES_VI_arroz_tartana-ES_VI_catedral → combinação do tipo (1)-(1)
- FR_AD_bric_a_br-FR_AD_femme → combinação do tipo (1)-(1)

[29] Teste T é um teste de hipótese que avalia se as médias de duas amostras A e B são significativamente diferentes. Ele considera que a probabilidade p representa a significância do resultado. Ou seja, p é parâmetro de julgamento (rejeição ou validação) da hipótese nula.

- ES_VI_arroz_tartana-FR_AD_bric_a_brac → combinação do tipo (1)-(2)
- ES_VI_arroz_tartana-FR_AD_femme → combinação do tipo (1)-(2)
- ES_VI_catedral-FR_AD_bric_a_brac → combinação do tipo (1)-(2)
- ES_VI_catedral-FR_AD_femme → combinação do tipo (1)-(2)

O procedimento descrito acima é realizado, da mesma forma, para as demais classes.

5. RESULTADOS E DISCUSSÕES

Este capítulo dedica-se à análise, caracterização e diferenciação das redes dos textos literários (originais e aleatórios) utilizados nesse trabalho, bem como ao teste da hipótese relacionada à formação de agrupamentos no espaço dos índices segundo os atributos de idioma, conteúdo e autor. A avaliação está fundamentada na existência de uma rede ótima de palavras, denominada rede crítica, que, acredita-se, expresse o maior número de informação significativa do texto com um mínimo de ruído (informações pouco significativas).

Para identificar-se o valor da Força-Fidelidade normalizada que está associada à rede crítica, analisou-se o comportamento de dois índices em função da Força-Fidelidade normalizada (FF_N): o caminho mínimo médio (CMM) e a diferença normalizada (ΔD_N) entre o número de vértices e o número de arestas da rede de palavras que compõe o texto. Estes índices foram escolhidos por apresentarem pontos que representam mudanças expressivas na topologia da rede. Os valores de Forças-Fidelidades correspondentes a tais pontos chamamos por Força-Fidelidade Crítica (*FFc*).

O caminho mínimo médio, como foi dito anteriormente, é um índice estatístico determinado pela média sobre todos os valores de caminhos mínimos da rede de associação de palavras, enquanto que a diferença normalizada entre o número de vértices e o número de arestas da rede assume uma intensidade para cada valor de Força-Fidelidade (variação ponto-a-ponto). Esta diferença normalizada (ΔD_N) é dada pela expressão

$$\Delta D_N = \frac{(V_i - V_{\min})}{(V_{\max} - V_{\min})} - \frac{(A_i - A_{\min})}{(A_{\max} - A_{\min})} \qquad \text{em que} \qquad (5.1)$$

- V_i representa o número de vértices da rede referente à Força-Fidelidade normalizada i
- V_{min} e V_{max} são, respectivamente, os números mínimos e máximos de vértices das diversas redes
- A_i representa o número de arestas da rede referente à Força-Fidelidade normalizada i

- A_{min} e A_{max} são, respectivamente, os números mínimos e máximos de arestas das diversas redes

5.1 IDENTIFICAÇÃO DAS FORÇAS-FIDELIDADES CRÍTICAS

5.1.1 ANÁLISE DOS TEXTOS ORIGINAIS

O intervalo adotado para gerar a Força-Fidelidade, nos textos originais, e subseqüente detecção do valor crítico, foi de 5×10^{-5} a 5×10^{-3}, com 200 pontos de análise. Ele foi estabelecido após uma verificação inicial do comportamento da rede para vários intervalos e quantidades de pontos de análises. Como produto desta primeira avaliação, foi possível encontrar um intervalo único em que todos os textos originais apresentassem pontos críticos. Este intervalo está explicitado acima.

Vale ressaltar que também foi analisado o comportamento da rede construída para $FF_N = 0$, ou seja, sem qualquer eliminação de arestas. Esta rede foi chamada de Rede Canônica.

Para cada valor de Força-Fidelidade normalizada (FF_N) gerado, existe uma rede de palavras correspondente. É a partir do valor mínimo adotado de FF_N que se executou a filtragem: arestas e vértices isolados foram eliminados permanecendo apenas associações cujos valores de FF_N são maiores ou iguais à intensidade de FF_N considerada.

É importante destacar que nesta subsecção, apresenta-se a análise gráfica dos diversos índices de rede para apenas um texto ou, quando possível, um texto para cada idioma, visto que todos apresentaram comportamentos semelhantes. Esta medida foi tomada com o fim de não tornar a discussão repetitiva e cansativa.

Tomando o comportamento do caminho mínimo médio (CMM) em função da Força-

Fidelidade normalizada (Figura 32), é possível notar um evidente ponto crítico que corresponde a um valor máximo bem pronunciado. Este comportamento crítico foi observado em todos os 50 textos literários avaliados, independentemente do idioma, com algumas variações na largura da curva.

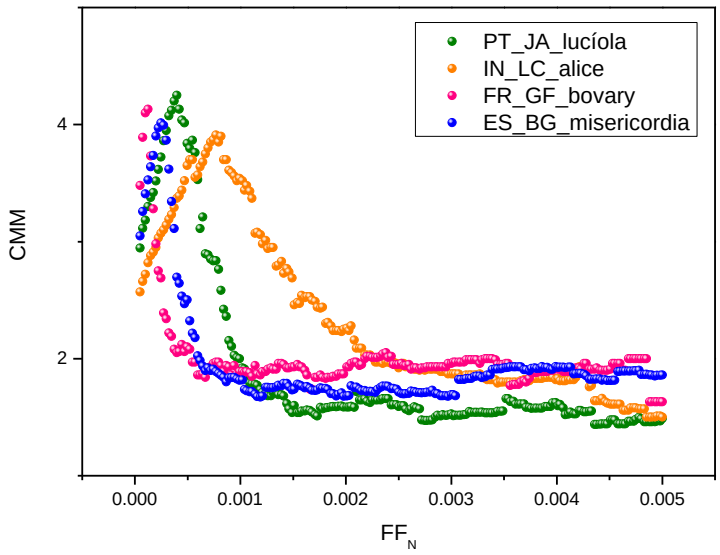


Figura 32. Representação gráfica do caminho mínimo médio em função da Força-Fidelidade normalizada para quatro textos literários de autores, conteúdos, idiomas e tamanhos (kb) diferentes
Fonte: Elaborado pela autora com base em dados da amostra, 2009

Como foi mencionado no capítulo 3, o caminho mínimo representa a menor distância existente entre dois vértices da rede. Neste caso, a menor distância entre duas palavras na rede semântica. Partindo do menor valor de FF_N adotado, o CMM apresenta um crescimento à medida que a Força-Fidelidade cresce. Esta resposta à filtragem modifica-se a partir de um determinado valor de FF_N. Este aumento do CMM se deve à perda dos atalhos presentes nestas redes. Note que a varredura promovida pelo aumento da Força-Fidelidade normalizada representa um ataque sistemático às arestas que conectam as palavras. Isso significa que a estrutura da rede demonstra a resistência das associações de palavras mais significativas a esses ataques.

Este fato fica melhor compreendido com a análise do número de vértices e arestas em função da FF_N. Observe na Figura 33 que, conforme a Força-Fidelidade normalizada

aumenta, inicialmente, o número de arestas decai mais rapidamente do que o número de vértices. Para intensidades maiores que certo valor de FF_N, há uma inversão e o número de vértices da rede passa a decrescer numa magnitude maior que o número de arestas. Por este valor de FF_N citado acima, fez-se passar uma reta tracejada afim de sinalizar o ponto de máximo do CMM expresso na Figura 32. Portanto, para valores de FF_N mais altos que aquele identificado, a rede vai ficando mais desconexa e mantendo cada vez menos associações e vocabulário, o que representa uma perda importante de informação (a rede vai sendo aos poucos "desmontada").

A interpretação para esse fenômeno é:

a) para valores de FF_N anteriores àquele em que ocorre o $CMM_{máx}$, isto é, o ponto crítico, tem-se a presença de muitas associações pouco significativas entre pares de palavras (muito ruído)

b) para valores de FF_N posteriores àquele em que ocorre o $CMM_{máx}$, tem-se a presença de poucas associações fortes entre pares de palavras (pouca informação)

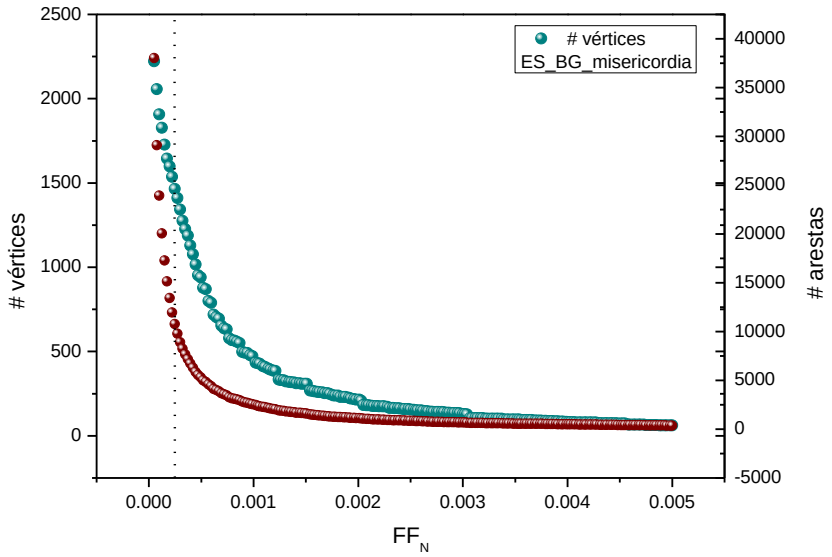


Figura 33. Representação do número de vértices e número de arestas em função da Força-Fidelidade normalizada para o texto ES_BG_misericordia
Fonte: Elaborado pela autora com base em dados da amostra, 2009

É importante salientar que, assim como no trabalho de Teixeira (2010), o comportamento observado para o CMM é promovido, fundamentalmente, pelo termo vindo da Fidelidade explícito na equação (3.8). Ou seja, ao se considerar apenas a força de interação entre pares de palavras, o CMM não apresentaria pontos críticos (Figura 34).

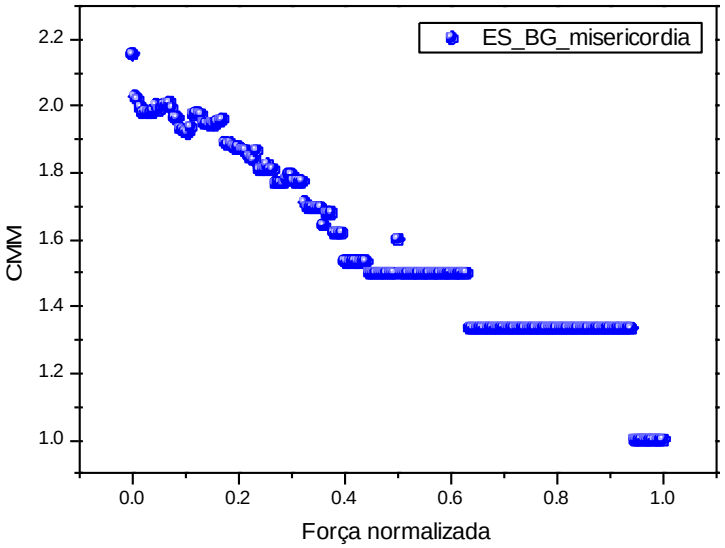


Figura 34. Comportamento do caminho mínimo médio da rede em função da Força normalizada para o texto ES_BG_misericordia
Fonte: Elaborado pela autora com base em dados da amostra, 2009

A diferença normalizada (ΔD_N)[30] existente entre o número de vértices e o número de arestas para cada Força-Fidelidade normalizada é representada pela Figura 35.

[30] O comportamento observado na Figura 35 se repete, com alguma variação na largura da curva, para todos os 50 textos analisados.

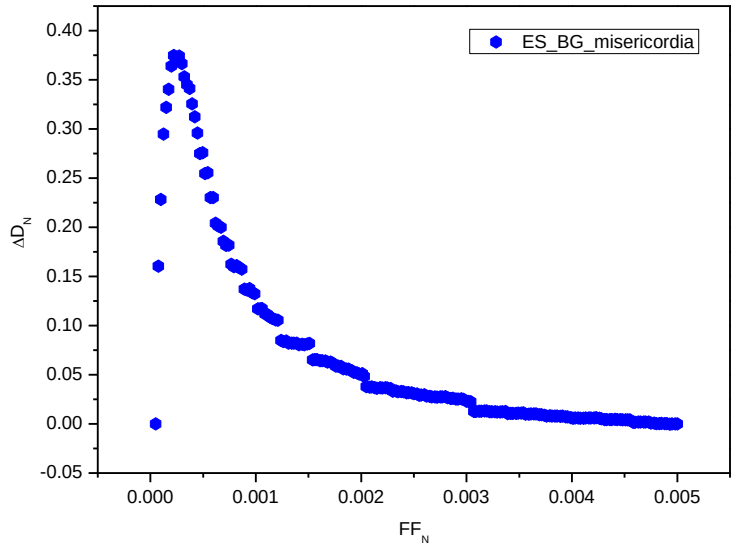


Figura 35. Representação do comportamento da diferença normalizada entre o número de vértices e número de arestas (ΔD_N) em função da Força-Fidelidade normalizada (FF_N) para o texto ES_BG_misericordia
Fonte: Elaborado pela autora com base em dados da amostra, 2009

Em média, os valores de FF_N correspondentes aos pontos máximos das curvas representadas pelas Figura 32 e Figura 35 são muito próximos. Sua diferença média é de 1.82 $\times$ 10^{-5}, isto é, menor que a diferença entre dois valores consecutivos de Força-Fidelidade normalizada.

Acredita-se que tanto o valor do ponto crítico correspondente ao $CMM_{máx}$ quanto o associado ao $\Delta D_{N\max}$ podem ser utilizados para identificar o valor da Força-Fidelidade Crítica (*FFc*). Como o $CMM_{máx}$ é uma medida clássica de redes complexas que representa uma característica global da rede (medida estatística da menor distância entre pares de palavras), foi adotado nesse trabalho como critério para a determinação da Força-Fidelidade Crítica (*FFc*) e, conseqüentemente, da rede crítica de palavras dos diversos textos literários. Com isso, a FF_N assume tripla função, são elas:

(a) peso — pois nossa rede de associação de palavras é uma rede ponderada;

(b) parâmetro de controle — pois, através da varredura deste valor dentro do intervalo de FF_N adotado, é possível determinar a rede crítica;

(c) índice — pois o valor de $FF_N = FFc$ será considerado um índice caracterizador da rede

crítica.

É importante destacar que se investigou a existência de alguma correlação entre o número de palavras da rede canônica com a *FFc* com o fim de se verificar possíveis dependências com a ordem dessa rede. No APÊNDICE B, apresenta-se uma tabela com algumas quantidades que caracterizam numericamente todos os textos originais examinados nessa pesquisa. Baseando-se nestes dados, representa-se graficamente a Força-Fidelidade crítica em função do número de vértices da rede canônica (Figura 36).

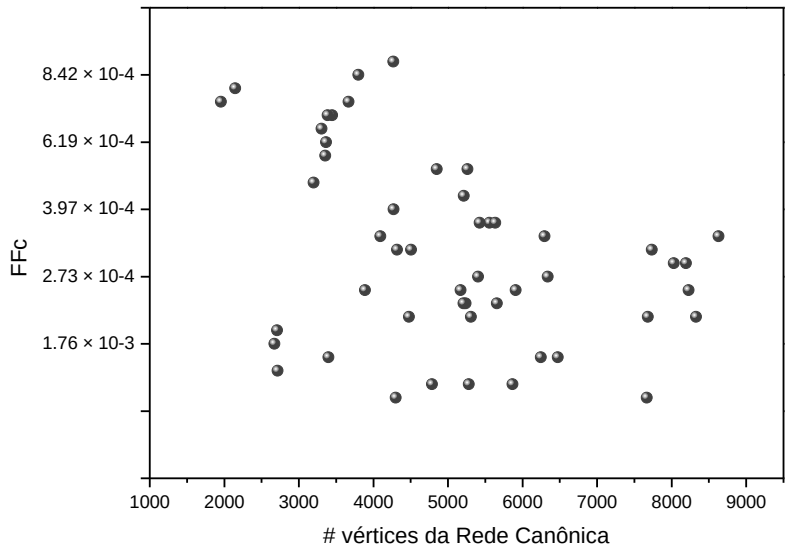


Figura 36. Representação gráfica da Força-Fidelidade Crítica (*FFc*) em função do número de vértices da Rede Canônica para cada um dos 50 textos analisados
Fonte: Elaborado pela autora com base em dados da amostra, 2009

Essa nuvem de pontos indica que existe uma flutuação muito grande sobre os valores referentes à *FFc*, de tal forma que este índice não depende da ordem da rede canônica[31]. O mesmo comportamento foi observado considerando a *FFc* em função da ordem da rede crítica.

[31] No APÊNDICE E, encontram-se as informações relacionadas às diversas redes canônicas dos textos originais.

Além disso, é aparente a tendência crescente[32] do número de vértices da rede crítica quando o número de vértices da rede canônica aumenta (Figura 37). Ou seja, quanto maior o número de vértices da rede canônica, maior será o número de vértices da rede crítica, porém não revelando uma dependência linear.

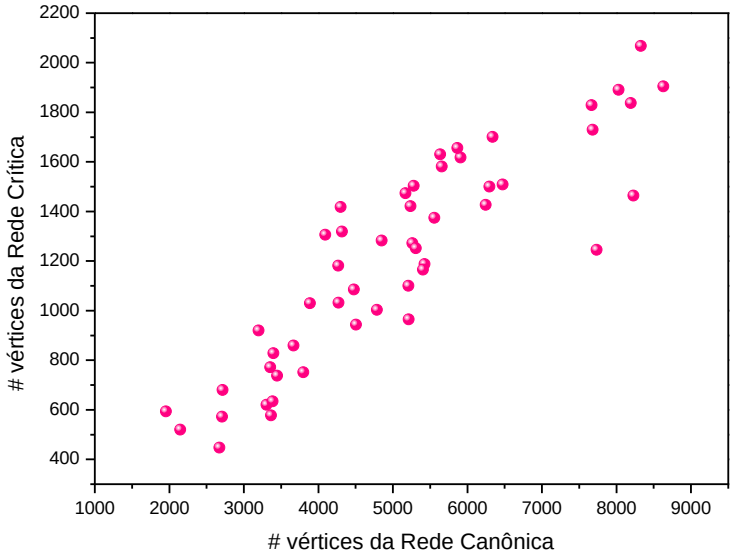


Figura 37. Representação do número de palavras da Rede Crítica em função do número de palavras da Rede Canônica para cada um dos 50 textos analisados
Fonte: Elaborado pela autora com base em dados da amostra, 2009

5.1.2 ANÁLISE DOS TEXTOS EMBARALHADOS

Considerando, nos textos embaralhados, a mesma quantidade de pontos e o mesmo intervalo de Força-Fidelidade que nos textos originais, observou-se que todos os índices avaliados se mantinham constantes. Isso significa que qualquer variação de FF_N sofrida neste intervalo não foi suficiente para promover alteração na estrutura da rede.

[32] Foi realizado o ajuste de uma reta na Figura 37. O coeficiente angular desta reta foi de, aproximadamente, 0.22(2) com R=0.89.

Assim, mantendo fixo o número de prospecções da Forças-Fidelidade em 200 pontos e modificando o intervalo escolhido para análise dos índices e redes de palavras oriundas dos textos embaralhados para 5×10^{-3} e 5×10^{-1}, obteve-se os seguintes resultados:

(i) **Caminho mínimo médio em função da Força-Fidelidade normalizada**

Diferentemente dos textos originais, 44% dos textos embaralhados não apresentaram pontos de máximo caminho mínimo médio. Isto significa que estes textos não possuem, segundo o critério adotado, um valor correspondente à Força-Fidelidade crítica.

Aproximadamente, 38% dos textos analisados possuem pontos de máximo bem definidos. Para essa classificação, foi estabelecido um critério observando-se a diferença entre os valores do caminho mínimo médio da rede inicial ($FF_N = 5 \times 10^{-3}$) e da rede crítica. Se essa diferença fosse maior do que um, então o ponto de máximo era considerado como um ponto crítico. Para ilustrar este comportamento, considere os quatro textos, que passaram por um processo de embaralhamento, representados na Figura 38. Destes, apenas o texto identificado pelo código RND_IN_LC_alice[33] apresenta um pico bem definido.

No APÊNDICE D encontra-se uma tabela contendo os índices característicos de redes complexa considerados nesta pesquisa para os 19 textos embaralhados.

[33] ATENÇÃO! Lembre-se que RND_ é usado para designar os textos embaralhados.

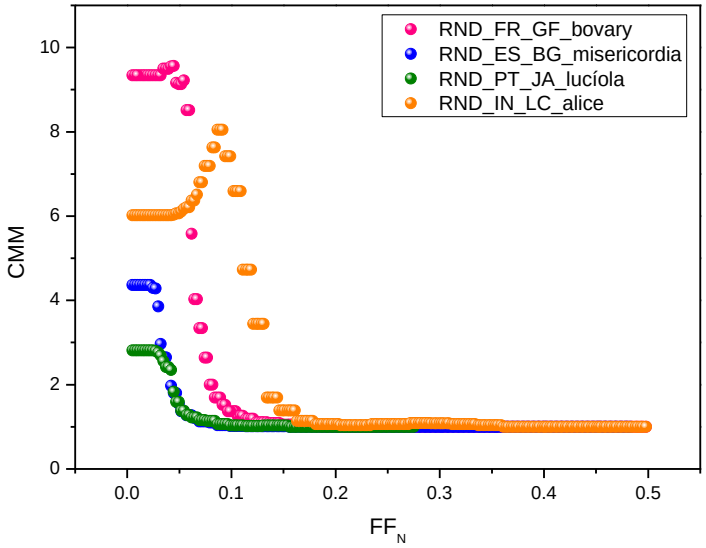


Figura 38. Representação gráfica do caminho mínimo médio em função da Força-Fidelidade normalizada para quatro textos embaralhados de autores, conteúdos, idiomas e tamanhos (kb) diferentes
Fonte: Elaborado pela autora com base em dados da amostra, 2009

É preciso deixar claro que, embora nem todos os textos tenham apresentado pontos de máximo, todos mostraram uma mudança de comportamento à medida que se varreu o intervalo adotado de FF_N. Resta entender, posteriormente, o porquê de alguns textos apresentarem tais pontos críticos e outros não.

Ao se comparar os resultados obtidos para os textos originais e os textos embaralhados listados nas tabelas dos APÊNDICES C e D, pode-se observar que os textos embaralhados possuem CMM maior que para os textos originais.

(ii) **Número de vértices e número de arestas em função da Força-Fidelidade normalizada**

Apesar das diferenças apontadas acima, todos os textos embaralhados, independentemente de possuírem pontos de máximo caminho mínimo médio, apresentaram comportamento semelhante para o número de vértices e o número de arestas em função da FF_N (Figura 39).

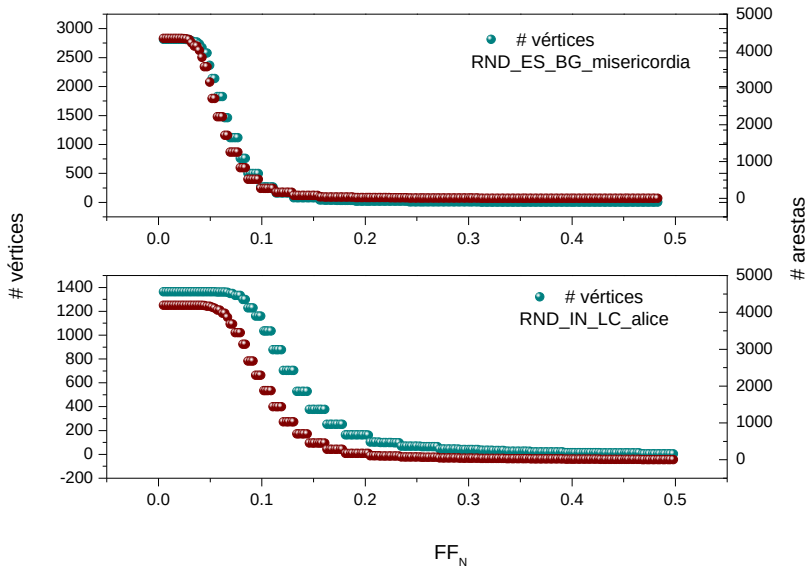


Figura 39. Representação do número de vértices e número de arestas em função da Força-Fidelidade normalizada para os textos RND_ES_BG_misericordia e RND_IN_LC_alice
Fonte: Elaborado pela autora com base em dados da amostra, 2009

Note que tanto os gráficos que representam o CMM (Figura 38) quanto o número de vértices e número de arestas (Figura 39) exibem um platô. Este platô informa que mesmo com os ataques sistemáticos, a estrutura da rede não é afetada.

Confrontando a Figura 33 e a Figura 39, pode-se também perceber diferenças no comportamento do número de vértices e número de arestas em relação à Força-Fidelidade normalizada. Tais evidências tornam-se mais claras com a análise realizada no item seguinte.

(iii) **Diferença normalizada entre o número de vértices e o número de arestas em função da Força-Fidelidade normalizada**

A diferença normalizada (ΔD_N) existente entre o número de vértices e o número de arestas para cada Força-Fidelidade normalizada (FF_N) é apresentada na Figura 40.

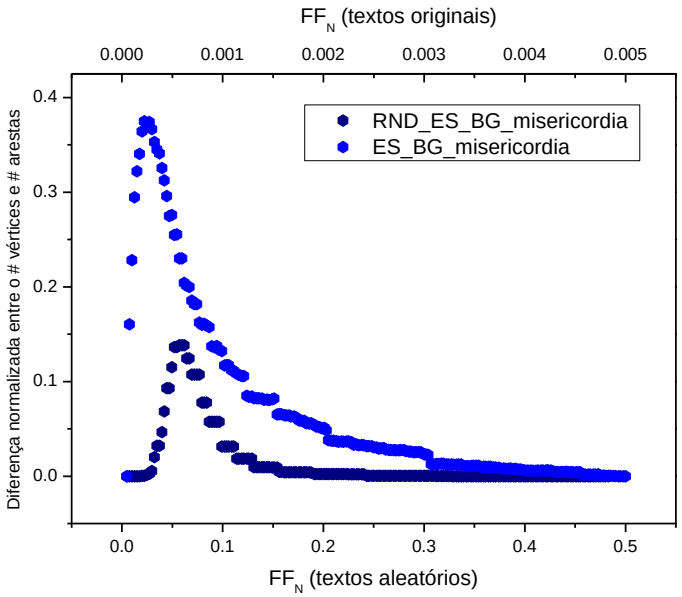


Figura 40. Representação gráfica para ΔD_N em função de FF_N para os textos ES_BG_misericordia (original) e RND_ES_BG_misericordia (aleatório)
Fonte: Elaborado pela autora com base em dados da amostra, 2009

Note que, segundo a Figura 40, tanto o texto original quanto embaralhado apresentam pontos de máximos. Porém, a amplitude deles e o valor de FF_N sobre o qual ΔD_N é máximo são muito diferentes.

Se fosse analisada apenas Figura 40, talvez não fosse suficiente para se inferir qualquer afirmação a respeito da capacidade que o método tem de captar diferenças entre cada uma das naturezas dos textos (textos originais e embaralhados). Porém, um conjunto de informações extraídas das figuras mostradas e uma análise da média dos índices de rede, mostrada na Tabela 4, parecem indicar que o método utilizado distingue um texto produzido por um indivíduo daquele que foi construído a partir de um processo mecânico de embaralhamento. Ou seja, a Força-Fidelidade parece ser um índice capaz de separar aquilo que faz parte da linguagem humana do que não faz parte.

A Tabela 4 representa um sumário de alguns índices de rede avaliados nesta pesquisa considerando somente os textos, originais e embaralhados, que apresentaram pontos de máximo.

NATUREZA DOS TEXTOS	*FFc*	D	CAM	CMM	<k>
Original	4.4×10^{-4}	14(2)*	0.20(4)	4.0(3)	6(2)
Embaralhado	6.3×10^{-2}	46(15)	0.04(1)	10(1)	1.9(2)

Tabela 4. Sumário contendo o valor médio para alguns dos índices de rede analisados
Fonte: Elaborado pela autora com base em dados da amostra, 2009
*Nesta pesquisa, foi calculado o Desvio Padrão

Para identificar a topologia dessas redes, faz-se necessária a análise da distribuição de graus. Esta etapa de caracterização das redes de palavras será realizada na secção seguinte.

5.2 CARACTERIZAÇÃO DAS REDES CRÍTICAS DOS TEXTOS ORIGINAIS E EMBARALHADOS

Para se classificar a topologia das redes de palavras oriundas dos textos analisados (originais e embaralhados), buscou-se verificar o comportamento da distribuição de graus. Através da análise dessa distribuição, é possível avaliar se a rede crítica de palavras oriundas dos textos originais é do mesmo tipo daquela oriunda dos textos embaralhados.

A Figura 41 representa, em log-log, uma distribuição de graus do tipo lei de potência com expoente 1.5(1) para o texto IN_LC_alice. Todas as distribuições de graus foram estudadas para o valor correspondente à *FFc*.

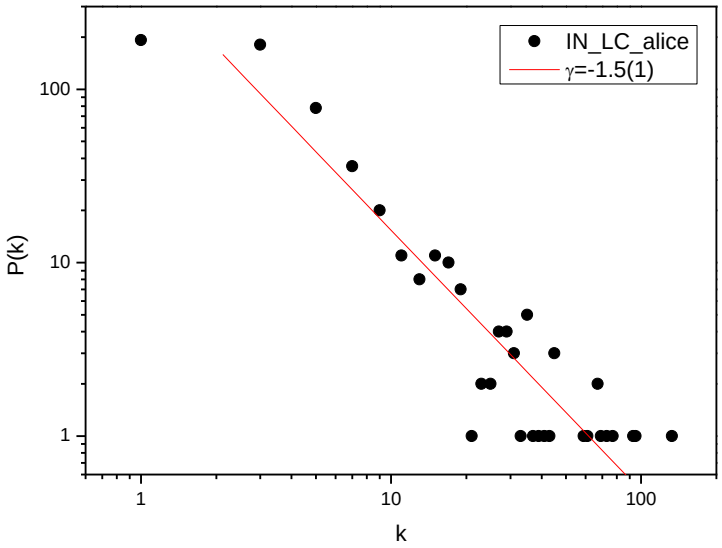


Figura 41. Distribuição de graus do tipo Lei de Potência para o texto IN_LC_alice
Fonte: Elaborado pela autora com base em dados da amostra, 2009

Este comportamento demonstra que as redes de palavras para textos originais exibem uma topologia do tipo livre de escala. Tal padrão foi identificado em todos os textos originais, independentemente do idioma. O expoente médio, obtido das 50 distribuições, é de 1.7(2).

A título de comparação, segue a Tabela 5 contendo a média dos índices de redes complexas extraídas desta pesquisa e dos trabalhos de Caldeira (2005) e Teixeira (2007).

TRABALHOS	# TEXTOS	D	CAM	CMM	γ
(1) Redes de textos escritos (2005)	312	5	0.77	2.3	1.6
(2) Redes de textos orais (2007)	12	4	0.80	2.1	1.7
(3) Redes de textos escritos	50	14	0.20	4.0	1.7

Tabela 5. Sumário contendo o valor médio aproximado para os índices de rede analisados, em três trabalhos distintos por ordem cronológica
Fonte: Elaborado pela autora, 2009

Ao se comparar os resultados desse trabalho com estes trabalhos anteriores, tem-se

a) os trabalhos envolvendo textos escritos ((1) e (3)) apresentam métodos diferentes de

construção da rede: o último constrói uma rede ponderada e calcula os índices característicos de rede para cada valor de *FF*. Isso significa que os índices médios acima apresentados são estabelecidos sobre condições diferentes.

Com base nesse argumento, buscou-se calcular a média aritmética sobre os 50 textos analisados para cada um desses índices. Essas médias foram determinadas considerando-se a rede canônica (que corresponde à rede do trabalho de Caldeira (2005)). Os valores médios de tais índices são: D=5, CAM= 0.74, CMM=2.3 e γ=1.8. Note que estas quantidades coincidem com os índices médios determinados por Caldeira (2005).

b) Os trabalhos envolvendo o mesmo método ((2) e (3)) apresentam três variáveis diferentes: o próprio processo de produção da linguagem, a quantidade de textos e o tamanho das redes (os discursos transcritos possuem tamanhos diferentes dos textos literários considerados nesse trabalho).

Analisando os dados apresentados nas tabelas do APÊNDICE C, comparou-se o comportamento do número de vértices da rede crítica dos diversos textos originais com o D, CAM, CMM e γ (Figura 42). Desta análise, não é possível identificar correlações entre as variáveis, o que se leva a crer que a diferença topológica existente entre os textos orais (2) e os textos escritos (3) não pode ser atribuído ao tamanho dos textos. Resta-se levantar a hipótese de que a diferença existente entre esses dois trabalhos é devido ao processo de produção da linguagem (oral x escrito).

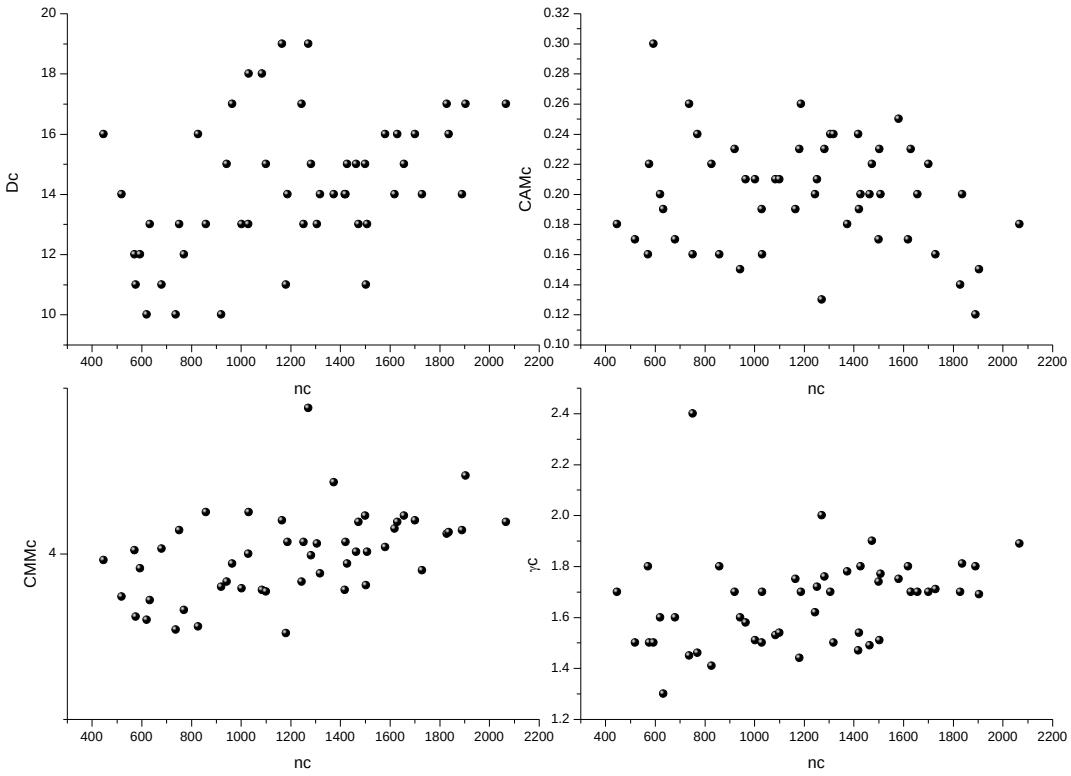


Figura 42. Análise do comportamento dos diversos números de vértices da rede crítica em função dos valores de D, CAM, CMM e γ extraídos também da rede crítica
Fonte: Elaborado pela autora com base em dados da amostra, 2009

Na Figura 43, apresenta-se quatro redes do texto Madame Bovary (em Francês) para quatro valores de Forças-Fidelidade normalizadas diferentes: (a) $FF_N = 0$ (Rede canônica), (b) $FF_N = 5 \times 10^{-5}$, (c) $FFc = 1.24 \times 10^{-4}$ (Rede crítica) e (d) $FFn = 5 \times 10^{-3}$.

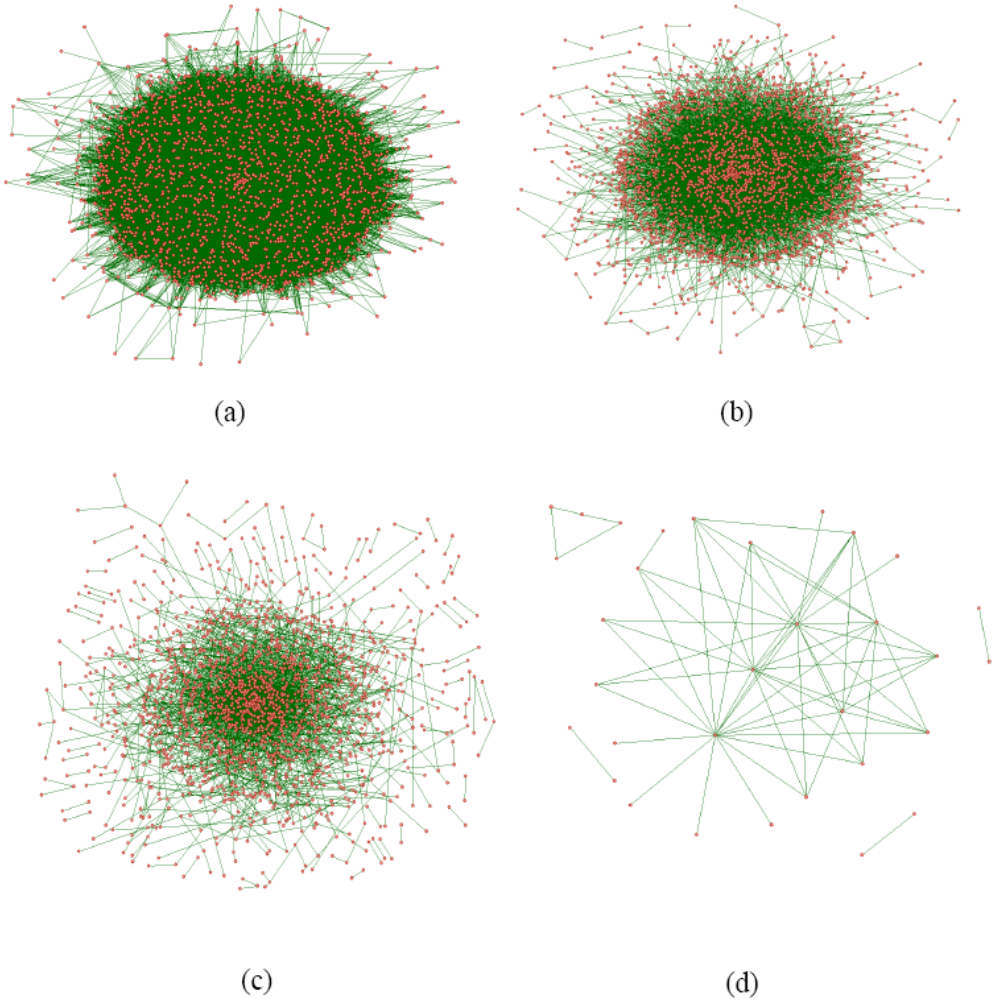


Figura 43. Ilustração de quatro redes de palavras que constituem o texto Madame Bovary, escrito em Francês, para quatro valores de FF_N distintas: (a) 0 (rede canônica), (b) 5×10^{-5}, (c) 1.24×10^{-4} (rede crítica) e (d) 5×10^{-3}
Fonte: Elaborado pela autora com base em dados da amostra, 2009

Com relação às redes de palavras oriundas dos textos embaralhados, foram avaliadas apenas as distribuições de graus para os 19 textos que apresentaram Força-Fidelidade Crítica. Em todos esses casos, foram identificadas características de uma rede do tipo aleatória.

A Figura 44 mostra que a distribuição de graus, em escala logarítmica, para um texto embaralhado não obedece a uma lei de potência e sim, apresenta uma distribuição do tipo normal.

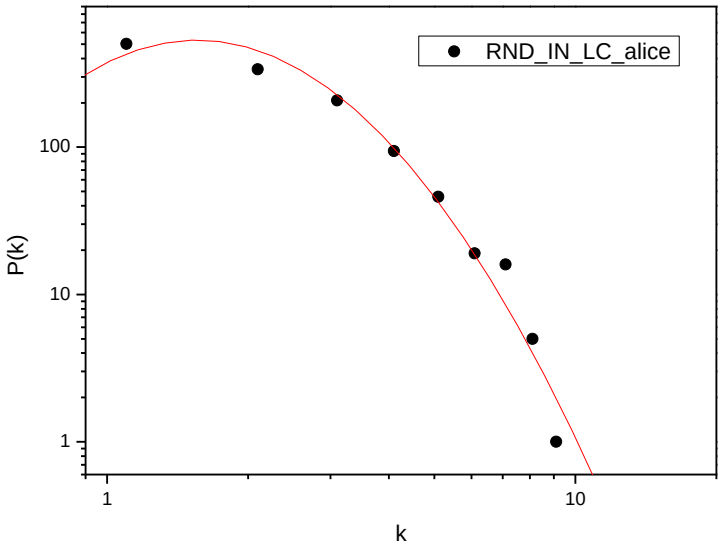


Figura 44. Distribuição de graus representada por uma parábola na escala di-log para o texto RND_IN_LC_alice
Fonte: Elaborado pela autora com base em dados da amostra, 2009

O comportamento expresso pela Figura 41 e Figura 44 indica que o método proposto por Teixeira (2007), e modificado durante a essa pesquisa, é sensível à forma de construção da rede.

Estas figuras e a Tabela 4 demonstram que a rede crítica de palavras oriunda de uma construção humana exibe uma topologia diferente daquela criada por um processo de embaralhamento realizado por um programa.

Na Figura 45 e Figura 46, respectivamente, mostra-se a rede crítica de palavras construída, originalmente, a partir do texto *Alice's Adventures in Wonderland* e após este texto ter sido submetido ao processo de embaralhamento.

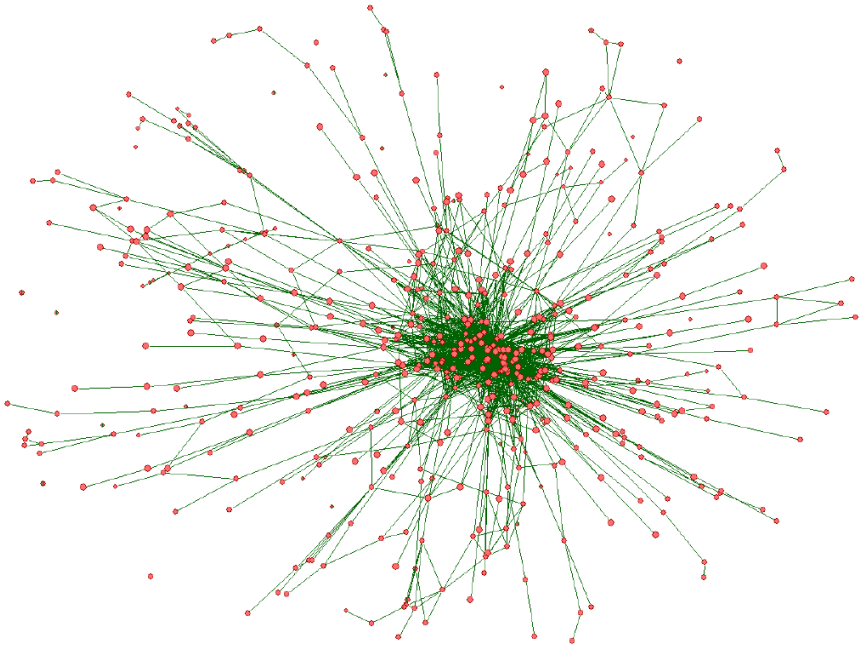


Figura 45. Representação, em 3D, da rede crítica de palavras oriundas do texto IN_LC_alice (texto original)
Fonte: Elaborado pela autora com base em dados da amostra, 2009

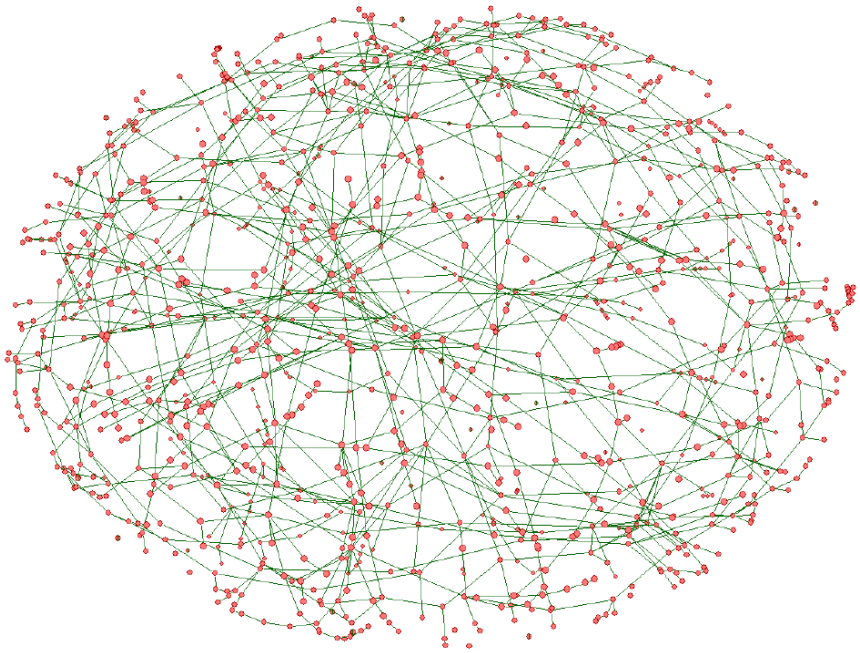


Figura 46. Representação, em 3D, da rede crítica de palavras oriundas do texto RND_IN_LC_alice (texto embaralhado)
Fonte: Elaborado pela autora com base em dados da amostra, 2009

5.3 TESTE DAS HIPÓTESES RELACIONDAS À FORMAÇÃO DE GRUPOS

De posse de todos os índices de redes relacionados ao valor de Força-Fidelidade crítica para todos os textos, selecionou-se 36 textos e agrupou-se, convenientemente, conforme cada classe.

As Tabela 6, Tabela 7 e Tabela 8. mostram os 12 textos selecionados para cada uma dessas classes e os índices da rede crítica adotados para análise.

NUMERAÇÃO DO GRUPO	TEXTO	*FFc*	D	CAM	CMM	<k>	γ
1	ES_VI_arroz_tartana	3.47×10^{-4}	17	0.15	4.47	4.9	1.69
1	ES_VI_catedral	2.98×10^{-4}	16	0.2	3.94	6.4	1.81
1	ES_VI_muertos	1.99×10^{-4}	17	0.18	4.19	5.54	1.89
2	FR_AD_bric_a_brac	3.72×10^{-4}	14	0.26	4.07	5.61	1.7
2	FR_AD_femme	5.20×10^{-4}	15	0.23	3.99	6.76	1.76
2	FR_AD_mille	3.47×10^{-4}	13	0.24	4.06	7.19	1.74
3	PT_MA_dom	1.49×10^{-4}	13	0.21	3.79	6.59	1.51
3	PT_MA_helena	3.22×10^{-4}	15	0.15	3.83	5.09	1.62
3	PT_MA_memorial	1.74×10^{-4}	16	0.22	3.56	7.49	1.41
4	IN_CD_chimes	6.44×10^{-4}	10	0.2	3.6	6.02	1.6
4	IN_CD_cricket	6.19×10^{-4}	11	0.22	3.62	5.34	1.5
4	IN_CD_house_let	5.95×10^{-4}	12	0.24	3.66	7.4	1.46

Tabela 6. Classe AUTOR e seus respectivos textos e índices críticos
Fonte: Elaborado pela autora com base em dados da amostra, 2009

NUMERAÇÃO DO GRUPO	TEXTO	*FFc*	D	CAM	CMM	<k>	γ
1	PT_JV_bovary	2.98×10^{-4}	14	0.12	4.14	5.73	1.76
1	FR_GF_bovary	1.24×10^{-4}	14	0.14	4.12	6.05	1.71
1	IN_GF_bovary	1.99×10^{-4}	14	0.16	3.9	6.57	1.71
2	PT_JV_balao	3.72×10^{-4}	14	0.18	4.43	4.53	1.78

NUMERAÇÃO DO GRUPO	TEXTO	*FFc*	D	CAM	CMM	<k>	γ
2	FR_JV_ballon	1.49×10^{-4}	15	0.2	4.23	6.76	1.76
2	IN_JV_balloon	2.73×10^{-4}	16	0.22	4.2	7.11	1.73
3	PT_JV_centro_terra	5.20×10^{-4}	19	0.13	4.88	3.3	1.98
3	FR_JV_centre_terre	1.74×10^{-4}	13	0.2	4.01	5.76	1.77
3	IN_JF_centre_earth	2.48×10^{-4}	14	0.17	4.15	6.16	1.81
4	PT_JV_volta	3.72×10^{-4}	16	0.23	4.19	8.58	1.7
4	IN_JV_around	2.48×10^{-4}	13	0.22	4.19	6.14	1.89
4	FR_JV_tour	2.23×10^{-4}	16	0.25	4.04	6.9	1.75

Tabela 7. Classe CONTEÚDO e seus respectivos textos e índices críticos
Fonte: Elaborado pela autora com base em dados da amostra, 2009

NUMERAÇÃO DO GRUPO	TEXTO	*FFc*	D	CAM	CMM	<k>	γ
1	ES_BG_marianela	4.71×10^{-4}	17	0.21	3.94	6.04	1.58
1	ES_VI_barraca	1.74×10^{-4}	15	0.2	3.94	5.74	1.82
1	ES_JV_pasarse	2.23×10^{-4}	15	0.21	3.77	6.99	1.54
2	FR_AD_capitaine	3.22×10^{-4}	14	0.24	3.88	10.45	1.5
2	FR_GA_don_juan	2.73×10^{-4}	19	0.19	4.2	5.01	1.75
2	FR_PF_loup_blanc	2.23×10^{-4}	14	0.19	4.07	8.26	1.54
3	IN_LC_alice	7.68×10^{-4}	12	0.3	3.91	6.31	1.5
3	IN_JA_love	4.96×10^{-4}	10	0.23	3.8	6.14	1.69
3	IN_DD_robinson	1.24×10^{-4}	14	0.24	3.78	9.91	1.47
4	PT_JA_viuvinha	1.860×10^{-3}	12	0.16	4.02	4.2	1.81
4	PT_MA_mao	6.69×10^{-4}	13	0.19	3.72	5.72	1.34
4	PT_AA_cortico	3.47×10^{-4}	15	0.17	4.23	6.49	1.74

Tabela 8. Classe IDIOMA e seus respectivos textos e índices críticos
Fonte: Elaborado pela autora com base em dados da amostra, 2009

É possível notar que foram considerados seis índices de rede para o cálculo da distância: cinco índices usuais (D, CAM, CMM, <k> e γ) e a Força-Fidelidade normalizada.

A partir dessas tabelas, calculou-se a distância entre os textos[34] e realizou-se o teste T, considerando um intervalo de confiança de 95%. Ou seja, $\alpha = 0.05$.

[34] Os resultados referentes ao cálculo da distância entre os textos estão apresentados nos apêndices F, G e H.

Com isso, as médias das distâncias entre os textos são significativamente diferentes se $p < 0.05$, caso contrário, são significativamente iguais.

A Tabela 9 apresenta, resumidamente, o resultado oriundo do Teste T.

	AUTOR		**CONTEÚDO**		**IDIOMA**	
	Intragrupo	Intergrupo	Intragrupo	Intergrupo	Intragrupo	Intergrupo
Média	0,7(2)	1,2(3)	0,8(5)	0,9(4)	0,9(3)	1,0(3)
p	< 0.001*		0,554		0,551	

Amostra: Intergrupo = 12, Entregrupo = 54 *Significativamente diferente

Tabela 9. Sumário do Teste T avaliando todas as classes analisadas nesta pesquisa
Fonte: Elaborado pela autora, 2009

Analisando esta tabela, pode-se perceber que:

(1) Existem diferenças significativas entre as médias das distâncias dos textos que compõem a classe autor. Isto indica que as redes de palavras dos textos literários para cada autor são, topologicamente, mais similares entre si do que se quando se compara estas redes de palavras com as redes oriundas dos textos de dois autores diferentes.

(2) Não é possível, a partir do método proposto, detectar diferenças topológicas significativas entre textos que compõem idioma[35]. Isto sugere que textos escritos em línguas diferentes parecem apresentar propriedades topológicas semelhantes (considerando o mesmo sistema de escrita).

(3) Quanto ao conteúdo, parece haver uma forte influência do tradutor na produção do texto, visto que apresentam distâncias significativamente iguais quando comparados aos textos pertencentes ao intragrupo e intergrupo. Ou seja, a distância entre dois textos com o mesmo conteúdo é significativamente igual à distância ente dois textos substancialmente diferentes.

[35] Este resultado corrobora com o trabalho de Ferrer i Cancho *et al*. (2004) e Caldeira (2006), por exemplo, e com a idéia de uniformidade da linguagem apresentada no Capítulo 2.

6. CONSIDERAÇÕES FINAIS

Neste último capítulo, apresentam-se tanto as conclusões extraídas desta pesquisa, quanto às perspectivas e possibilidades de investigações que surgem de alguns questionamentos postos a partir das análises e discussões dos resultados deste trabalho.

6.1 CONCLUSÕES

Neste trabalho, foram avaliados alguns aspectos da linguagem verbal escrita sob o ponto de vista da Teoria de Redes Complexas, a partir da análise de um conjunto de 50 textos literários escritos. Para isso, modelou-se computacionalmente[36] a linguagem escrita utilizando a idéia de rede semântica e o conceito de Força-Fidelidade proposto por Teixeira (2007).

Na primeira parte desta pesquisa, demonstrou-se, a partir de análise comparativa e quantitativa, que as redes de textos escritos originais apresentam topologias diferentes das redes de textos embaralhados. Ou seja, o método se mostrou eficiente na detecção de diferentes maneiras de construção dos textos (textos originais e aqueles que passaram por um processo de embaralhamento).

Com isso, pode-se notar que, invariavelmente, todas as redes de textos originais analisadas exibiram um comportamento crítico bem definido. Contudo, apenas a detecção dessa característica não é suficiente para promover a diferenciação desses dois tipos de rede, visto que a existência dessa mudança de comportamento também é observada em 38% dos textos embaralhados.

Isto é, a análise dos índices de rede e distribuição de graus é fundamental para

[36] O tempo estimado para o processamento dos textos depende não só das características da máquina, mas também do tamanho do texto em Kb, Forças-Fidelidades limites (FF_L) e número de pontos considerados neste intervalo de FF_L. Assim, para uma máquina Intel(R) Pentium(R) Dual CPU E2180 @ 2.0GHz com 1GB de memória RAM e adotando o intervalo de Força-Fidelidade utilizado nesta pesquisa, o tempo médio de processamento de cada texto foi de, aproximadamente, 47 min.

diferenciar as redes críticas de textos originais em livres de escala, também encontrada em trabalhos anteriores (FERRER I CANCHO (2001), CALDEIRA (2005), CORSO *et al* (2006), TEIXEIRA (2007)), daquelas redes críticas de textos embaralhados (redes aleatórias). Porém, a característica de mundo pequeno, comum a estes trabalhos anteriores, não foi observada.

Além disso, o método parece capturar diferenças entre as redes oriundas de textos orais e escritos revelando que, possivelmente, seus processos de produção sejam distintos.

Na segunda parte, demonstrou-se que a estrutura topológica da rede de palavras dos textos literários revela mais significativamente as características do autor do que as demais classes. Isto nos leva a sugerir o método como uma proposta para testes de reconhecimento de autor.

6.2 PERSPECTIVAS

A princípio, trabalhos de pesquisa de caráter interdisciplinar podem significar dificuldade, mas acabam por admitir investigações em diversas áreas do conhecimento. Ou seja, podem ser sinônimo de “solo fértil”.

Considerando esse aspecto, o presente estudo permite certa diversidade de abordagens. Na área da Lingüística, por exemplo, podem-se realizar análises comparativas entre outros estilos literários. Em computação, promover o desenvolvimento de softwares capazes de realizar buscas semânticas. E, sob o ponto de vista da “pesquisa de fronteira”, na qual a Física está inserida, pode-se:

- analisar, comparativamente, a distância euclidiana entre os textos da rede canônica e da rede crítica;
- analisar o comportamento do produto da Força-Fidelidade normalizada pela freqüência de ocorrência da palavra;
- analisar a distância entre redes a partir do método proposto por Andrade *et al*. (2008);

- realizar análise de campo médio, buscando entender mais profundamente o comportamento das redes dos textos embaralhados;
- verificar padrões de modularidade da rede de palavras;
- realizar análise utilizando o conceito de percolação;
- analisar, com a mesma base de dados, outros índices de redes, tais como assortatividade, *betweenness*, dimensão fractal e outros;
- expandir a análise utilizando uma base de dados maior, tornando os resultados estatisticamente relevantes;
- comparar efeitos de produção da fala e da escrita.

Além disso, pode-se sugerir também sua utilização como forma de avaliação de cunho educativo (avaliação de conhecimento adquirido, criatividade, comparação entre saberes, dentre outros).

REFERÊNCIAS

ALBERT, Réka, BARABÁSI, A.lbert-László. Statistical mechanics of complex networks. **Reviews of Modern Physics**, v. 74, jan. 2002, p. 47-97.

AMARAL, L.A.N, OTTINO J.M. Complex networks: Augmenting the framework for the study of complex systems. **Eur. Phys. J. B**, v. 38, 2004, p. 147–162.

ANDRADE, R. F. S., MIRANDA, J. G. V., PETIT LOBÃO, T. Neighborhood properties of complex network. **Phys. Rev.E**, v. 73, 2006., 046101.

ANTÔNIO, Juliano Desiderato. Diferenças lingüísticas produzidas por diferenças de modalidade de língua e de tipologia textual. **Guairacá**, v. 17, 2001, p. 7-22.

ANTIQUEIRA, L. *et al*. Strong correlations between text quality and complex networks features. **Physica A**, v. 373, 2007, p. 811–820.

BENTO, Conceição Aparecida. A escrita e o sujeito: uma leitura à luz de Lacan. Psicologia v. 15, n. 1-2, USP, 2004, p. 195-214. Disponível em: <http://www.scielo.br/pdf/pusp/v15n1-2/a20v1512.pdf>. Acessado em 10.11.2006.

BOCCALETTI, S. et al. Complex networks: Structure and dynamics. **Physics Reports**. v. 424, 2006, p. 175-308.

BORDENAVE, Juan E. Diaz. **Além dos meios e mensagens**: Introdução à comunicação como processo, tecnologia, sistema e ciência. 6. ed. Petrópolis: Vozes, 1993.

BURIANOVÁ, Zuzana. Do tempo na narrativa ao tempo em primeiras estórias. **Romanica Olomucensia VIII**, Olomouc, Vydavatelství UP Olomouc (CZE). 1999, v. 74, p. 19-30. Disponível em <http://publib.upol.cz/~obd/fulltext/Romanica-8/Romanica-8_03.pdf> Acessado em 04.06.2007.

CALDEIRA, Sílvia G M. **Caracterização da rede de signos lingüísticos:** um modelo baseado no aparelho psíquico de Freud. 2005. 131 f. Dissertação de Mestrado — Centro de Pesquisa e Pós-Graduação da Faculdade Visconde de Cairu (CEPPEV), Salvador, 2005.

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

CANCHO, Ramon Ferrer I., SOLÉ, Ricard V. The small world of human language. **Proc. R. Soc. London B**, v. 268, 2001, p. 2261-2265.

CANCHO, R. F. I, SOLÉ, R. V., KÖHLER, R. Patterns in syntactic dependency networks. **PHYSICAL REVIEW E**, v. 69, 051915, 2004.

CORRÊA, R. H. M. A. Literatura, texto e hipertexto. **Terra roxa e outras terras: Revista de Estudos Literários**, v. 8, 2006, p. 30-43. ISSN 1678-2054. Disponível em: <http://www.uel.br/cch/pos/letras/terraroxa>. Acessado em 09.10.2007.

CORSO, G. *et al*. A Scale-free Network of Evoked Words. **Brazilian Journal of Physics**, 2006, v. 36, n. 3A, set. 2006. Disponível em: <http://redalyc.uaemex.mx/redalyc/pdf/464/46436432.pdf>. Acessado em 02.03.2008.

COSTA, Luciano da Fontoura. Redes "Complexas": modelagem 'simples' da natureza. **Ciência Hoje**, v. 36, n. 216, mar. 2005, p. 34-39.

CUNHA, Dóris de Arruda Carneiro. Visitando a interação na prosa literária. **D.E.L.T.A.**, v. 24, n. 1, p. 105-123, 2008, ISSN 0102-4450. Disponível em: <http://www.scielo.br/cgi-bin/wxis.exe/iah/>. Acessado em 12.01.2009.

DOROGOVTSEV, S. N., MENDES, J. F. F. Language as an evolving word web. **Proc. R. Soc. Lond. B**, v. 268, 2001, p. 2603-2606.

DIAS, M. C. P. Cognição e modelos computacionais: Duas abordagens. **Veredas - Revista de Estudos Lingüísticos**. v. 4, n. 1, 2000, Juiz de Fora, p. 31-41. ISSN 1415-2533.

EYSENCK, Michael W., KEANE, Mark T. **Psicologia Cognitiva**: um Manual Introdutório. Tradução Wagner Gesser e Maria Helena Fenalti Gesser. Porto Alegre: Artes Médicas, 1994.

FANTI, Maria da Glória Corrêa Di. A linguagem em Bakthin: pontos e despontos. **Veredas - Revista de Estudos Lingüísticos**. v. 7, n. 1 e 2, Juiz de Fora, 2003., p. 95-111.

FONSECA, Suzana Carielo. Lesão x sintoma: uma questão sobre a causalidade. **DELTA**. São Paulo, v. 14, n. 12, 1998. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44501998000200008>. Acessado em 23.10.2006.

FREUD, Sigmund. **A interpretação das Afasias**. Tradução de Antonio Pinho Ribeiro. 1. ed. Lisboa: Edições 70, 1979.

GALVÃO, Viviane Matos. **Um modelo para neoplasia utilizando redes complexas**. 2006. 105 f. Dissertação de Mestrado – Programa de Pós-Graduação do Instituto de Física da Universidade Federal da Bahia (UFBA), Salvador, 2006.

GAZZANIGA, M. S. *et al*. **Neurociência Cognitiva**: a biologia da mente. 2. ed. Porto Alegre: Artmed Bookman, 2006. p. 369-416.

GRIFFITHS T. L.; STEYVERS M.; TENENBAUM J. B. Topics in Semantic Representation. **Psychological Review**, 2007, v. 114, n. 2, p. 211-244.

GRZYBEK, Peter; KÖHLER, Reinhard. **Exact Methods in the study of language and text**. Berlim: Mounton de Gruyter, 2007. Disponível em:

<http://books.google.com.br/books?id=EziPylQdXycC&printsec=frontcover&dq=GRZYBEK>. Acessado em 21.02.2009

FERES JÚNIOR, João. De Cambridge para o mundo, historicamente: Revendo a contribuição metodológica de Quentin Skinner. **DADOS – Revista de Ciências Sociais**, Rio de Janeiro, v. 48, n. 3, 2005, p. 655-680. Disponível em: <http://www.scielo.br/pdf/dados/v48n3/a07v48n3.pdf>. Acessado em 03.07.2008.

KOLB, B., WHISHAW, I.Q. **Neurociência do comportamento**. São Paulo: Manole, 2002.

LOFFREDO, A.M. Em busca do referente, às voltas com a polissemia dos sonhos: a questão em Freud, Stuart Mill e Lacan. **Psicologia USP**, São Paulo, v. 10, n. 1, 1999, p. 169-97. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-65641999000100009&lng=en&nrm=iso> Acessado em 24.10.2006.

LUBISCO, N. M. L., VIEIRA S. C., SANTANA, I. V. **Manual de estilo acadêmico**: monografias, dissertações e teses. 3. ed. Salvador: EDUFBA, 2007.

MANIS, Melvin. **Processos Cognitivos**. Tradução de Olgierd Ligeza-Stamiriwski. São Paulo: Helder, 1973. 221 f. Coleção Ciência do Comportamento.

MATTA, Isabel. Aprender vivendo: As experiências de vida no desenvolvimento e na aprendizagem. **Análise Psicológica**. v. 22, n. 1, 2004, p. 73-80.

MIZRAJI, Eduardo, VALLE-LISBOA, Juan C. **Schizophenic speech as disordered trajectory in a collapsed cognitive "Small-World"**. Medical Hypotheses, v. 68, p. 347-352, 2007 (In impress)

NEWMAN, M. E. J.; STROGATZ, S. H.; WATTS D. J. Random graphs with arbitrary degree distributions and their applications. **Physical Review E**, v. 64, 2001, 026118.

NEWMAN, M. E. J. The Structure and Function of Complex Networks. **SIAM Review**. v. 45, n. 2, 2003, p. 167-256.

ORLANDI, Eni Pulcinelli. **O que é Lingüística**. São Paulo: Brasiliense, 1999. 71 f. Coleção Primeiros Passos; 184.

PEREIRA. Mirna Feitoza. Contenha-se, se for capaz. **Galáxia**. n. 4, 2002. p. 263-269. Disponível em: <http://revistas.pucsp.br/index.php/galaxia/article/viewFile/1297/794>. Acessado em 02.01.2009.

PETRONI, Maria Rosa. **LINGUASAGEM**: Revista Eletrônica de Popularização Científica em Ciências da Linguagem. ISSN: 1983-6988. Disponível em: <http://www.letras.ufscar.br/linguasagem/edicao01/materialdidatico_generosdodiscurso.htm>. Acessado em 10.03.2009.

PINHO, S. T. R. **Modelo de Ising em redes aperiódicas e criticalidade auto-organizada**.

Tese de Doutorado, Instituto de Física - Universidade de São Paulo, São Paulo, 1998, p. 43-54.

QUEIROZ, João. Novos modelos de cognição encorporada, situada e contextualizada em Ciências Cognitivas. **Revista Eletrônica Informação e Cognição**, v. 2, n. 1, 2000, p. 37-43,. ISSN: 1807-8281

QUEIROZ, Rita de C. R. de. A informação escrita: do manuscrito ao texto Virtual. **Diálogo Científica**, dez. 2005. Disponível em: <http://dici.ibict.br/archive/00000513/>. Acessado em 30.11.2006.

SANTANA, Charles N. **Análise da Pluviometria do Nordeste Brasileiro segundo modelagem em redes**. Monografia. UFBA - Universidade Federal da Bahia. 2005.

SASSURE, Ferdinand. **Curso de Lingüística Geral**. 27 ed. São Paulo: Cultrix, 2006. P. 79-82.

SCLIAR-CABRAL, Leonor. Inter-relações entre a componente semântica e memória episódica. **Veredas - Revista de Estudos Lingüísticos**. v. 6, n. 1, 2002, Juiz de Fora, p. 105-112. ISSN 1415-2533

SOUZA, Débora de Hollanda. Falando sobre a Mente: Algumas considerações sobre a relação entre a linguagem e a teoria da mente. **Psicologia: Reflexão e Crítica**, v. 19, n. 3, 2005, p. 387-394. Disponível em: <http://www.scielo.br/pdf/prc/v19n3/a07v19n3.pdf>. Acessado em 06.10.2006.

STEYVERS, Mark; TENENBAUM Joshua B. The Large-Scale Structure of Semantic Networks:Statistical Analyses and a Model of Semantic Growth. **Cognitive Science**. v. 29, 2005, p. 41-78.

SZCZESNIAK, Konrad. Palavras-relâmpago: Como aprendemos e utilizamos nosso vocabulário. **Ciência Hoje**, v. 35, n.207, ago. 2004, p. 17-20.

TELES, Gilberto Mendonça. Para uma poética do conto brasileiro. **Revista de Filologia Românica**. v. 19, 2002, p. 161-182.

TEIXEIRA, Gesiane M. **Redes Semânticas em discursos orais**: Uma proposta metodológica baseada na psicologia cognitiva utilizando redes complexas. 2007. 118 f. Dissertação de Mestrado - Centro de Pesquisa e Pós-Graduação da Faculdade Visconde de Cairu (CEPPEV), Salvador, 2007.

TEIXEIRA, G. M. *et al*. Complex semantic networks. **International Journal of Modern Physics C**. v. 21, 2010, p. 333-347.

UNITEX: Manual de Utilização. Tradução Alexis Neme e Oto Araújo Vale. Rede Relex Brasil: 2002, cap. 0-4. Disponível em: <http://ladl.univ-mlv.fr/brasil/Ferramentas/Unitex.html>. Acessado em 16.09.2006.

UNITEX 1.2: User Manual. Sébastien Paumier: jun. 2006. Disponível em: <http://www-igm.univ-mlv.fr/~unitex/UnitexManual.pdf>. Acessado em 13.01.2007.

VASCONCELLOS, Zilda. A frase do ponto de vista semântico. **Cadernos do Congresso Nacional de Linguística e Filologia**. v. 11, n. 11, Círculo Fluminense de Estudos Filológicos e Lingüísticos. Rio de Janeiro, 2008.

APÊNDICES

APÊNDICE A: Composição dos textos literários analisados e suas classificações

CÓDIGO	PUBLICAÇÃO	AUTOR	IDIOMAS	TÍTULO	GÊNERO LITERÁRIO	# PALAVRAS	#FRASES	VOCABULÁRIO
ES_BG_marianela	1878	Benito Pérez Galdós (1843-1920)	Espanhol	Marianela	Romance	50993	3058	5214
ES_BG_misericordia	1897	Benito Pérez Galdós (1843-1920)	Espanhol	Misericórdia	Romance	83877	4222	8232
ES_BG_torquemada	1888	Benito Pérez Galdós (1843-1920)	Espanhol	Torquemada em la Hoguera	Romance	62402	3263	7749
ES_JV_pasarse_listo	1906	Juan Valera (1824-1905)	Espanhol	Pasarse de Listo	Romance	50929	2844	5236
ES_VI_arroz_tartana	1894	Vicente Blasco Ibánes (1867-1928)	Espanhol	Arroz y Tartana	Romance	91015	4868	8619
ES_VI_barraca	1898	Vicente Blasco Ibánes (1867-1928)	Espanhol	La Barraca	Romance	54299	2848	6263
ES_VI_catedral	1903	Vicente Blasco Ibánes (1867-1928)	Espanhol	La Catedral	Romance	98128	5522	8210
ES_VI_muertos	1909	Vicente Blasco Ibánes (1867-1928)	Espanhol	Los Muertos Mandan	Romance	98380	5673	8345
FR_AD_bric_a_brac	1861	Alexandre Dumas (1802-1870)	Francês	Bric-a-Brac	Conto	49775	2485	5497
FR_AD_captaine	1838	Alexandre Dumas (1802-1870)	Francês	Le Capitaine Paul	Romance	63595	2278	4307
FR_AD_femme	1851	Alexandre Dumas (1802-1870)	Francês	La Femme au Collier de Velours	Romance	57536	2441	4891
FR_AD_mille	1849	Alexandre Dumas (1802-1870)	Francês	Les Mille et un Fantômes	Conto	55825	2607	4262

CÓDIGO	PUBLICAÇÃO	AUTOR	IDIOMAS	TÍTULO	GÊNERO LITERÁRIO	# PALAVRAS	#FRASES	VOCABULÁRIO
FR_EA_homme	1862	Edmond About (1828-1885)	Francês	L'Homme à l'Oreille Cassée	Romance	58218	2974	5326
FR_GA_don_juan	1914	Guillaume Apollinaire (1880-1918)	Francês	Les Trois Don Juan	Romance	58656	3562	5424
FR_GF_bovary	1857	Gustave Flaubert (1821-1880)	Francês	Madame Bovary	Romance	116176	6389	7662
FR_JA_notre	1896	Jean Aicard (1848-1921)	Francês	Notre-Dame-D'Amour	Romance	48626	2701	3916
FR_JV_ballon	1863	Jules Verne (1828-1905)	Francês	Cinq Semaines en Ballon	Romance	82346	4469	5847
FR_JV_centre_terre	1864	Jules Verne (1828-1905)	Francês	Voyage au Centre de la Terre	Romance	69028	4005	6513
FR_JV_tour	1873	Jules Verne (1828-1905)	Francês	Le Tour du Monde en Quatre-vingts Jours	Romance	70082	3781	6300
FR_PF_fee-greves	1850	Paul Féval (Pére) (1816-1887)	Francês	La fée de Grèves	Romance	78257	4821	5543
FR_PF_loup_blanc	1843	Paul Féval (Pére) (1816-1887)	Francês	Le Loup Blanc	Romance	78171	4260	5238
IN_CD_chimes	1844	Charles Dickens (1812-1870)	Inglês	The Chimes	Novela	31689	2475	3301
IN_CD_cricket	1845	Charles Dickens (1812-1870)	Inglês	The Cricket on the Hearth	Novela	32551	2317	3361
IN_CD_house_let	1858	Charles Dickens (1812-1870)	Inglês	A House to Let	Conto	35031	2078	3350
IN_CD_haunted	1848	Charles Dickens (1812-1870)	Inglês	The Haunted Man and the Ghost's Bargain	Romance	34722	1401	3429
IN_DD_robinson	1719	Daniel Defoe (1660-1731)	Inglês	Robinson Crusoe	Romance	100976	3890	4304

CÓDIGO	PUBLICAÇÃO	AUTOR	IDIOMAS	TÍTULO	GÊNERO LITERÁRIO	# PALAVRAS	#FRASES	VOCABULÁRIO
IN_GF_bovary	-	Gustave Flaubert (1821-1880)	Inglês	Madame Bovary	-	117911	6864	7659
IN_JA_lady_susan	1871	Jane Austen (1775-1817)	Inglês	Lady Susan	Conto	23383	1378	2186
IN_JA_love	1822	Jane Austen (1775-1817)	Inglês	Love and Friendship	Conto	33808	1772	3186
IN_JA_northanger	1818	Jane Austen (1775-1817)	Inglês	Northanger Abbey	Romance	78692	4093	4462
IN_JA_persuasion	1818	Jane Austen (1775-1817)	Inglês	Persuasion	Romance	84166	4517	4503
IN_JV_around	-	Jules Verne (1828-1905)	Inglês	Around the World in Eighty Days	-	63905	3403	5150
IN_JV_balloon	-	Jules Verne (1828-1905)	Inglês	Five Weeks in a Balloon	-	92706	4419	6302
IN_JV_centre_earth	-	Jules Verne (1828-1905)	Inglês	A Journey to the Centre of the Earth	-	86283	4816	5901
IN_LC_alice	1865	Lewis Carroll (1832-1898)	Inglês	Alice's Adventures in Wonderland	Romance	27386	1684	1922
PT_AA_cortico	1890	Aluísio de Azevedo (1857-1913)	Português	O Cortiço	Romance	79445	4500	6321
PT_GF_bovary	-	Gustave Flaubert (1821-1880)	Português	Madame Bovary	-	113538	7352	8059
PT_JA_diva	1864	José de Alencar (1829-1877)	Português	Diva	Romance	33818	3002	3833
PT_JA_gazela	1870	José de Alencar (1829-1877)	Português	A Pata da Gazela	Romance	35193	2698	3714
PT_JA_iracema	1865	José de Alencar (1829-1877)	Português	Iracema	Romance	25043	1638	2746

CÓDIGO	PUBLICAÇÃO	AUTOR	IDIOMAS	TÍTULO	GÊNERO LITERÁRIO	# PALAVRAS	#FRASES	VOCABULÁRIO
PT_JA_luciola	1862	José de Alencar (1829-1877)	Português	Lucíola	Romance	45888	3020	4295
PT_JA_viuvinha	1857	José de Alencar (1829-1877)	Português	A Viuvinha	Romance	18113	1041	2724
PT_JV_balao	-	Jules Verne (1828-1905)	Português	Cinco Semanas em Balão	-	67655	4792	5563
PT_JV_centro_terra	-	Jules Verne (1828-1905)	Português	Viagem ao Centro da Terra	-	61374	5121	5310
PT_JV_volta	-	Jules Verne (1828-1905)	Português	A Volta ao Mundo em Oitenta Dias	-	65173	2633	5644
PT_MA_alienista	1881	Machado de Assis (1839-1908)	Português	O Alienista	Conto	16947	966	2675
PT_MA_dom	1899	Machado de Assis (1839-1908)	Português	Dom Casmurro	Romance	66884	4518	4789
PT_MA_helena	1876	Machado de Assis (1839-1908)	Português	Helena	Romance	56715	4255	4515
PT_MA_mao	1874	Machado de Assis (1839-1908)	Português	A Mão e a Luva	Romance	35266	2120	3390
PT_MA_memorial	1908	Machado de Assis (1839-1908)	Português	Memorial de Aires	Romance	51258	3417	3399

APÊNDICE B: Quantidades correspondentes ao tamanho (kbytes) e o número de palavras do texto (sem qualquer tratamento), bem como o vocabulário da rede canônica, Força-Fidelidade Crítica e o vocabulário referente a este valor para todos os textos literários que foram analisados (ordem crescente do tamanho em kb)

TEXTOS	TAMANHO (KB)	# PALAVRAS	VOCABULÁRIO DA REDE CANÔNICA	FF_C	VOCABULÁRIO DA REDE CRÍTICA
PT_MA_alienista	102	16947	2673	1.76×10^{-3}	447
IN_JA_lady_susan	128	23383	2149	8.17×10^{-4}	520
PT_JA_viuvinha	137	18113	2709	1.86×10^{-3}	572
PT_JA_iracema	145	25043	2714	1.51×10^{-3}	680
IN_LC_alice	151	27386	1956	7.68×10^{-4}	593
IN_CD_chimes	171	31689	3307	6.44×10^{-4}	620
IN_CD_cricket	179	32551	3365	6.19×10^{-4}	577
IN_CD_house_let	186	35031	3355	5.95×10^{-4}	771
IN_JA_love	186	33808	3198	4.96×10^{-4}	920
IN_CD_haunted	188	34722	3447	6.69×10^{-4}	737
PT_JA_diva	193	33818	3799	8.42×10^{-4}	751
PT_MA_mao	202	35266	3388	6.69×10^{-4}	634
PT_JA_gazela	203	35193	3667	7.68×10^{-4}	859
FR_JA_notre	276	48626	3887	2.48×10^{-4}	1030
PT_JA_lucíola	280	45888	4272	3.97×10^{-4}	1032
ES_JV_pasarse_listo	286	50929	5210	2.23×10^{-4}	1100
FR_AD_bric_a_brac	289	49775	5425	3.72×10^{-4}	1188
PT_MA_memorial	294	51258	3399	1.74×10^{-4}	828
ES_BG_marianela	294	50993	5213	4.71×10^{-4}	965

TEXTOS	TAMANHO (KB)	# PALAVRAS	VOCABULÁRIO DA REDE CANÔNICA	FF_C	VOCABULÁRIO DA REDE CRÍTICA
FR_AD_mille	317	55825	4095	3.47×10^{-4}	1306
ES_VI_barraca	320	54299	6247	1.74×10^{-4}	1427
PT_MA_helena	329	56715	5308	3.22×10^{-4}	943
FR_EA_homme	329	58218	4505	1.99×10^{-4}	1252
FR_AD_femme	330	57536	4849	5.20×10^{-4}	1283
FR_GA_don_juan	349	58656	5406	2.73×10^{-4}	1165
ES_BG_torquemada	359	62402	7734	3.22×10^{-4}	1245
FR_AD_capitaine	362	63595	4316	3.22×10^{-4}	1319
IN_JV_around	368	63905	5170	2.48×10^{-4}	1473
PT_MA_dom	370	66884	4788	1.49×10^{-4}	1003
PT_JV_centro_terra	374	61374	5262	5.20×10^{-4}	1272
PT_JV_volta	395	65173	5637	3.72×10^{-4}	1630
PT_JV_balao	408	67655	5556	3.72×10^{-4}	1374
FR_JV_centre_terre	418	69028	6475	1.74×10^{-4}	1509
FR_JV_tour	422	70082	5655	2.23×10^{-4}	1581
IN_JA_northanger	437	78692	4478	1.99×10^{-4}	1085
FR_PF_loup_blanc	449	78171	6295	2.23×10^{-4}	1421
FR_PF_fee_greves	451	78257	5236	1.49×10^{-4}	1503
PT_AA_cortico	465	79445	4267	3.47×10^{-4}	1500
IN_JA_persuasion	466	84166	8228	9.95×10^{-5}	1181
ES_BG_misericordia	478	83877	5908	2.48×10^{-4}	1464
IN_JV_centre_earth	485	86283	5868	2.48×10^{-4}	1618

TEXTOS	TAMANHO (KB)	# PALAVRAS	VOCABULÁRIO DA REDE CANÔNICA	FF_C	VOCABULÁRIO DA REDE CRÍTICA
FR_JV_ballon	490	82346	4300	1.49×10^{-4}	1656
IN_DD_robinson	527	100976	6339	1.24×10^{-4}	1418
IN_JV_balloon	528	92706	8628	2.73×10^{-4}	1701
ES_VI_arroz_tartana	536	91015	8194	3.47×10^{-4}	1905
ES_VI_catedral	572	98128	8329	2.98×10^{-4}	1837
ES_VI_muertos	581	98380	7683	1.99×10^{-4}	2068
IN_GF_bovary	652	117911	7666	1.99×10^{-4}	1730
FR_GF_bovary	684	116176	8029	1.24×10^{-4}	1829
PT_GF_bovary	684	113538	6295	2.98×10^{-4}	1891

APÊNDICE C: Valores de Forças-Fidelidades Críticas e índices característicos associados a estes valores para os diversos textos literários escritos em Espanhol, Francês, Inglês e Português

TEXTOS - ESPANHOL	FF_C	**n**	**m**	**D**	**CAM**	**CMM**	**<k>**	**Γ**
ES_BG_marianela	4.71×10^{-4}	965	5824	17	0.21	3.94	6.04	1.58(9)
ES_BG_misericordia	2.48×10^{-4}	1464	10760	15	0.20	4.01	7.35	1.49(8)
ES_BG_torquemada	3.22×10^{-4}	1245	6922	17	0.20	3.83	5.56	1.62(8)
ES_JV_pasarse_listo	2.23×10^{-4}	1100	7694	15	0.21	3.77	6.99	1.54(8)
ES_VI_arroz_tartana	3.47×10^{-4}	1905	9344	17	0.15	4.47	4.90	1.69(9)
ES_VI_barraca	1.74×10^{-4}	1427	8188	15	0.20	3.94	5.74	1.8(1)
ES_VI_catedral	2.98×10^{-4}	1837	11756	16	0.20	4.13	6.40	1.81(9)
ES_VI_muertos	1.99×10^{-4}	2068	11448	17	0.18	4.19	5.54	1.89(8)

TEXTOS - FRANCÊS	FF_C	**n**	**m**	**D**	**CAM**	**CMM**	**<k>**	**Γ**
FR_AD_bric_a_brac	3.72×10^{-4}	1188	6662	14	0.26	4.07	5.61	1.7(1)
FR_AD_capitaine	3.22×10^{-4}	1319	13786	14	0.24	3.88	10.45	1.50(8)
FR_AD_femme	5.20×10^{-4}	1283	8674	15	0.23	3.99	6.76	1.76(8)
FR_AD_mille	3.47×10^{-4}	1306	9386	13	0.24	4.06	7.19	1.7(1)
FR_EA_homme	1.99×10^{-4}	1252	6656	13	0.21	4.07	5.32	1.72(8)
FR_GA_don_juan	2.73×10^{-4}	1165	5832	19	0.19	4.20	5.01	1.75(8)
FR_GF_bovary	1.24×10^{-4}	1829	11072	17	0.14	4.12	6.05	1.7(1)
FR_JA_notre	2.48×10^{-4}	1030	6584	13	0.19	4.00	6.39	1.5(1)
FR_JV_ballon	1.49×10^{-4}	1656	11192	15	0.20	4.23	6.76	1.7(1)
FR_JV_centre_terre	1.74×10^{-4}	1509	8692	13	0.20	4.01	5.76	1.77(9)
FR_JV_tour	2.23×10^{-4}	1581	10906	16	0.25	4.04	6.90	1.75(8)
FR_PF_fee_greves	1.49×10^{-4}	1503	12202	11	0.23	3.81	8.12	1.51(9)
FR_PF_loup_blanc	2.23×10^{-4}	1421	11744	14	0.19	4.07	8.26	1.54(8)

TEXTOS - INGLÊS	FF_C	**n**	**m**	**D**	**CAM**	**CMM**	**<k>**	**Γ**
IN_CD_chimes	6.44×10^{-4}	620	3734	10	0.20	3.60	6.02	1.60(8)
IN_CD_cricket	6.19×10^{-4}	577	3082	11	0.22	3.62	5.34	1.50(8)
IN_CD_haunted	6.69×10^{-4}	737	6858	10	0.26	3.54	9.30	1.45(8)
IN_CD_house_let	5.95×10^{-4}	771	5706	12	0.24	3.66	7.40	1.46(9)
IN_DD_robinson	1.24×10^{-4}	1418	14058	14	0.24	3.78	9.91	1.47(8)
IN_GF_bovary	1.99×10^{-4}	1730	11364	14	0.16	3.90	6.57	1.71(8)
IN_JA_lady_susan	8.17×10^{-4}	520	2504	14	0.17	3.74	4.82	1.5(2)

TEXTOS - INGLÊS	FF_C	**n**	**m**	**D**	**CAM**	**CMM**	**<k>**	**γ**
IN_JA_love	4.96×10^{-4}	920	5648	10	0.23	3.80	6.14	1.7(1)
IN_JA_northanger	1.99×10^{-4}	1085	8594	18	0.21	3.78	7.92	1.53(9)
IN_JA_persuasion	9.95×10^{-5}	1181	13826	11	0.23	3.52	11.71	1.44(8)
IN_JV_around	2.48×10^{-4}	1473	9044	13	0.22	4.19	6.14	1.9(1)
IN_JV_balloon	2.73×10^{-4}	1701	12092	16	0.22	4.20	7.11	1.7(1)
IN_JV_centre_earth	2.48×10^{-4}	1618	9966	14	0.17	4.15	6.16	1.8(1)
IN_LC_alice	7.68×10^{-4}	593	3742	12	0.30	3.91	6.31	1.5(1)

TEXTOS - PORTUGUÊS	FF_C	**n**	**m**	**D**	**CAM**	**CMM**	**<k>**	**γ**
PT_AA_cortico	3.47×10^{-4}	1500	9738	15	0.17	4.23	6.49	1.74(9)
PT_GF_bovary	2.98×10^{-4}	1891	10842	14	0.12	4.14	5.73	1.8(1)
PT_JA_diva	8.42×10^{-4}	751	2382	13	0.16	4.14	3.17	2.4(2)
PT_JA_gazela	7.68×10^{-4}	859	3628	13	0.16	4.25	4.22	1.8(2)
PT_JA_iracema	1.51×10^{-3}	680	3630	11	0.17	4.03	5.34	1.6(1)
PT_JA_lucíola	3.97×10^{-4}	1032	4368	18	0.16	4.25	4.23	1.7(1)
PT_JA_viuvinha	1.86×10^{-3}	572	2400	12	0.16	4.02	4.20	1.8(2)
PT_JV_balao	3.72×10^{-4}	1374	6228	14	0.18	4.43	4.53	1.78(9)
PT_JV_centro_terra	5.20×10^{-4}	1272	4204	19	0.13	4.88	3.30	2.0(2)
PT_JV_volta	3.72×10^{-4}	1630	13984	16	0.23	4.19	8.58	1.70(8)
PT_MA_alienista	1.76×10^{-3}	447	1544	16	0.18	3.96	3.42	1.7(2)
PT_MA_dom	1.49×10^{-4}	1003	6608	13	0.21	3.79	6.59	1.51(9)
PT_MA_helena	3.22×10^{-4}	943	4800	15	0.15	3.83	5.09	1.6(1)
PT_MA_mao	6.69×10^{-4}	634	3628	13	0.19	3.72	5.72	1.3(1)
PT_MA_memorial	1.74×10^{-4}	828	6202	16	0.22	3.56	7.49	1.41(9)

APÊNDICE D: Valores de Forças-Fidelidades Críticas e seus correspondentes índices característicos para os 19 textos que passaram por um processo de embaralhamento e apresentaram pontos críticos bem definidos

TEXTOS	FF_C	**n**	**m**	**D**	**CAM**	**CMM**	**<k>**
RND_ES_JV_pasarse_listo	4.21×10^{-2}	2598	4876	36	0.04	9.39	1.88
RND_FR_AD_capitaine	3.47×10^{-2}	2822	6390	25	0.04	8.00	2.26
RND_FR_AD_femme	3.22×10^{-2}	2582	4818	35	0.04	8.50	1.87
RND_FR_AD_mille	5.20×10^{-2}	2137	3748	57	0.04	9.83	1.75
RND_FR_EA_homme	3.72×10^{-2}	2149	3578	77	0.03	10.4	1.66
RND_FR_JA_notre	8.17×10^{-2}	2014	3916	33	0.05	8.7	1.94
RND_FR_JV_ballon	7.68×10^{-2}	2851	4796	79	0.02	13.34	1.68
RND_FR_PF_fee_greves	2.73×10^{-2}	2696	4806	50	0.02	9.71	1.78
RND_FR_PF_loup_blanc	6.69×10^{-2}	2991	5624	42	0.03	11.36	1.88
RND_IN_CD_haunted	3.97×10^{-2}	1834	3604	31	0.06	9.31	1.97
RND_IN_DD_robinson	9.16×10^{-2}	2720	4932	49	0.01	11.27	1.81
RND_IN_JA_love	5.45×10^{-2}	1519	2912	43	0.06	8.98	1.92
RND_IN_JA_northanger	7.68×10^{-2}	2619	4810	38	0.02	10.16	1.84
RND_IN_JA_persuasion	1.09×10^{-1}	2871	5352	46	0.02	10.91	1.86
RND_IN_JV_around	7.93×10^{-2}	2365	4010	59	0.03	10.60	1.70
RND_IN_JV_balloon	7.18×10^{-2}	3391	6136	54	0.03	10.23	1.81
RND_IN_LC_alice	9.16×10^{-2}	1227	2678	24	0.03	8.05	2.18
RND_PT_JV_volta	5.7×10^{-2}	3030	5882	36	0.05	8.72	1.94
RND_PT_MA_mao	3.22×10^{-2}	1294	2138	57	0.05	8.00	1.65

APÊNDICE E: Relação dos índices característicos associados à Rede Canônica para os diversos textos literários escritos em Espanhol, Francês, Inglês e Português

TEXTO - ESPANHOL	**n**	**m**	**D**	**CAM**	**CMM**	**<k>**	**g**
ES_BG_marianela.txt_FF	5213	307674	5	0.78	2.17	59.02	1.97
ES_BG_misericordia.txt_FF	8228	561540	5	0.78	2.12	68.25	2.02
ES_BG_torquemada.txt_FF	7734	469950	4	0.78	2.12	60.76	2.02
ES_JV_pasarse.txt_FF	5210	314120	4	0.78	2.12	60.29	1.93
ES_VI_arroz_tartana.txt_FF	8628	646914	5	0.74	2.12	74.98	1.85
ES_VI_barraca.txt_FF	6247	383896	4	0.74	2.21	61.45	1.89
ES_VI_catedral.txt_FF	8194	543672	5	0.73	2.30	66.35	1.86
ES_VI_muertos.txt_FF	8329	567180	5	0.71	2.27	68.10	1.76

TEXTO - FRANCÊS	**n**	**m**	**D**	**CAM**	**CMM**	**<k>**	**g**
FR_AD_bric_a_brac.txt_FF	5425	311386	5	0.78	2.23	57.40	1.87
FR_AD_capitaine.txt_FF	4316	408230	5	0.75	2.13	94.59	1.64
FR_AD_femme.txt_FF	4849	348156	5	0.75	2.18	71.80	1.83
FR_AD_mille.txt_FF	4095	289172	5	0.74	2.19	70.62	1.75
FR_EA_homme.txt_FF	5308	304948	5	0.74	2.27	57.45	1.87
FR_GA_don_juan.txt_FF	5406	290118	5	0.72	2.31	53.67	1.92
FR_GF_bovary.txt_FF	7666	573218	4	0.71	2.25	74.77	1.82
FR_JA_notre.txt_FF	3887	240756	5	0.73	2.23	61.94	1.78
FR_JV_ballon.txt_FF	5868	394888	5	0.72	2.32	67.30	1.77
FR_JV_centre_terre.txt_FF	6475	372576	5	0.75	2.27	57.54	1.87
FR_JV_tour.txt_FF	5655	432058	4	0.75	2.16	76.40	1.72
FR_PF_fee_greves.txt_FF	5280	346124	5	0.73	2.27	65.55	1.80
FR_PF_loup_blanc.txt_FF	5236	382994	5	0.73	2.22	73.15	1.83

TEXTO - INGLÊS	**n**	**m**	**D**	**CAM**	**CMM**	**<k>**	**g**
IN_CD_chimes.txt_FF	3307	145476	5	0.77	2.37	43.99	1.77
IN_CD_cricket.txt_FF	3365	161924	5	0.77	2.28	48.12	1.80
IN_CD_haunted.txt_FF	3447	198776	4	0.78	2.21	57.67	1.84
IN_CD_house_let.txt_FF	3355	168738	5	0.76	2.28	50.29	1.78
IN_DD_robinson.txt_FF	4300	358218	4	0.73	2.11	83.31	1.65
IN_GF_bovary.txt_FF	7683	553170	5	0.71	2.31	72.00	1.77
IN_JA_lady_susan.txt_FF	2149	98030	4	0.73	2.16	45.62	1.74
IN_JA_love.txt_FF	3198	187118	5	0.72	2.20	58.51	1.71
IN_JA_northanger.txt_FF	4478	333690	4	0.72	2.17	74.52	1.71
IN_JA_persuasion.txt_FF	4267	338040	4	0.74	2.11	79.22	1.66
IN_JV_around.txt_FF	5170	297028	5	0.70	2.29	57.45	1.84
IN_JV_balloon.txt_FF	6339	480872	5	0.71	2.27	75.86	1.77
IN_JV_centre_earth.txt_FF	5908	399874	5	0.71	2.28	67.68	1.75
IN_LC_alice.txt_FF	1956	168168	5	0.76	2.05	85.98	1.49

TEXTO - PORTUGUÊS	n	m	D	CAM	CMM	<k>	g
PT_AA_cortico.txt_FF	6295	417716	5	0.71	2.38	66.36	1.82
PT_GF_bovary.txt_FF	8029	541868	5	0.70	2.40	67.49	1.83
PT_JA_diva.txt_FF	3799	116146	5	0.71	2.55	30.57	1.91
PT_JA_gazela.txt_FF	3667	131008	5	0.71	2.48	35.73	1.87
PT_JA_iracema.txt_FF	2714	102968	5	0.73	2.49	37.94	1.63
PT_JA_lucíola.txt_FF	4272	193976	5	0.71	2.41	45.41	1.93
PT_JA_viuvinha.txt_FF	2709	94628	5	0.76	2.49	34.93	1.80
PT_JV_balao.txt_FF	5556	278846	5	0.69	2.46	50.19	1.83
PT_JV_centro_terra.txt_FF	5262	225654	7	0.68	2.52	42.88	1.84
PT_JV_volta.txt_FF	5637	484484	5	0.74	2.23	85.95	1.68
PT_MA_alienista.txt_FF	2673	88290	5	0.77	2.41	33.03	1.88
PT_MA_dom.txt_FF	4788	229500	5	0.75	2.33	47.93	1.79
PT_MA_helena.txt_FF	4505	190278	5	0.71	2.40	42.24	1.91
PT_MA_mao.txt_FF	3388	149550	5	0.75	2.32	44.14	1.82
PT_MA_memorial.txt_FF	3399	155184	5	0.74	2.32	45.66	1.68

APÊNDICE F[37]: Resultados do cálculo da distância entre textos para a classe AUTOR

PARES DE TEXTOS *i* E *j*	$\delta_{i,j}$	**PARES DE TEXTOS *i* E *j***	$\delta_{i,j}$
ES_VI_arroz-ES_VI_catedral	0.986886	FR_AD_bric-PT_MA_helena	1.08086
ES_VI_arroz-ES_VI_muertos	0.702226	FR_AD_bric-PT_MA_memorial	1.25692
ES_VI_arroz-FR_AD_bric	1.20624	FR_AD_bric-IN_CD_chimes	1.12304
ES_VI_arroz-FR_AD_femme	1.24416	FR_AD_bric-IN_CD_cricket	0.99695
ES_VI_arroz-FR_AD_mille	1.41122	FR_AD_bric-IN_CD_house	1.11727
ES_VI_arroz-PT_MA_dom	1.3816	FR_AD_femme-FR_AD_mille	0.497253
ES_VI_arroz-PT_MA_helena	0.778112	FR_AD_femme-PT_MA_dom	1.00016
ES_VI_arroz-PT_MA_memorial	1.69935	FR_AD_femme-PT_MA_helena	1.10483
ES_VI_arroz-IN_CD_chimes	1.64402	FR_AD_femme-PT_MA_memorial	1.16261
ES_VI_arroz-IN_CD_cricket	1.58102	FR_AD_femme-IN_CD_chimes	1.01181
ES_VI_arroz-IN_CD_house	1.83963	FR_AD_femme-IN_CD_cricket	1.06495
ES_VI_catedral-ES_VI_muertos	0.554051	FR_AD_femme-IN_CD_house	0.893357
ES_VI_catedral-FR_AD_bric	0.753302	FR_AD_mille-PT_MA_dom	0.778255
ES_VI_catedral-FR_AD_femme	0.573687	FR_AD_mille-PT_MA_helena	1.23992
ES_VI_catedral-FR_AD_mille	0.676311	FR_AD_mille-PT_MA_memorial	1.06153
ES_VI_catedral-PT_MA_dom	0.840073	FR_AD_mille-IN_CD_chimes	1.10479
ES_VI_catedral-PT_MA_helena	0.810242	FR_AD_mille-IN_CD_cricket	1.1877
ES_VI_catedral-PT_MA_memorial	1.06852	FR_AD_mille-IN_CD_house	0.900824
ES_VI_catedral-IN_CD_chimes	1.25531	PT_MA_dom-PT_MA_helena	0.944015
ES_VI_catedral-IN_CD_cricket	1.29308	PT_MA_dom-PT_MA_memorial	0.650017
ES_VI_catedral-IN_CD_house	1.26261	PT_MA_dom-IN_CD_chimes	1.14853
ES_VI_muertos-FR_AD_bric	1.00476	PT_MA_dom-IN_CD_cricket	1.12236
ES_VI_muertos-FR_AD_femme	1.02582	PT_MA_dom-IN_CD_house	1.01768
ES_VI_muertos-FR_AD_mille	1.11237	PT_MA_helena-PT_MA_memorial	1.28564
ES_VI_muertos-PT_MA_dom	1.18128	PT_MA_helena-IN_CD_chimes	1.15522
ES_VI_muertos-PT_MA_helena	0.849031	PT_MA_helena-IN_CD_cricket	1.10298
ES_VI_muertos-PT_MA_memorial	1.48369	PT_MA_helena-IN_CD_house	1.4487
ES_VI_muertos-IN_CD_chimes	1.63125	PT_MA_memorial-IN_CD_chimes	1.4663
ES_VI_muertos-IN_CD_cricket	1.62644	PT_MA_memorial-IN_CD_cricket	1.43073
ES_VI_muertos-IN_CD_house	1.76215	PT_MA_memorial-IN_CD_house	1.05218
FR_AD_bric-FR_AD_femme	0.63615	IN_CD_chimes-IN_CD_cricket	0.410894
FR_AD_bric-FR_AD_mille	0.659719	IN_CD_chimes-IN_CD_house	0.772642
FR_AD_bric-PT_MA_dom	0.908022	IN_CD_cricket-IN_CD_house	0.835046

[37]Por uma questão operacional, os códigos de alguns textos presentes nos apêndices F, G e H foram reduzidos ou modificados. Porém, acredita-se que a maneira como foi exposto não compromete a compreensão da análise.

APÊNDICE G: Resultados do cálculo da distância entre textos para a classe CONTEÚDO

PARES DE TEXTOS *i* E *j*	$\delta_{i,j}$	**PARES DE TEXTOS *i* E *j***	$\delta_{i,j}$
PT_JV_bovary-FR_GF_bovary	0.502706	PT_JV_balao-FR_JV_centro	0.735298
PT_JV_bovary-IN_GF_bovary	0.523782	PT_JV_balao-IN_JF_centro	0.540725
PT_JV_bovary-PT_JV_balao	0.626307	PT_JV_balao-PT_JV_volta	0.994488
PT_JV_bovary-FR_JV_balao	0.771063	PT_JV_balao-IN_JV_volta	0.726494
PT_JV_bovary-IN_JV_balao	0.889017	PT_JV_balao-FR_JV_volta	0.95601
PT_JV_bovary-PT_JV_centro	1.5537	FR_JV_balao-IN_JV_balao	0.407809
PT_JV_bovary-FR_JV_centro	0.723486	FR_JV_balao-PT_JV_centro	1.76034
PT_JV_bovary-IN_JF_centro	0.449995	FR_JV_balao-FR_JV_centro	0.450153
PT_JV_bovary-PT_JV_volta	1.0963	FR_JV_balao-IN_JF_centro	0.441585
PT_JV_bovary-IN_JV_volta	0.927162	FR_JV_balao-PT_JV_volta	0.751365
PT_JV_bovary-FR_JV_volta	1.09899	FR_JV_balao-IN_JV_volta	0.654443
FR_GF_bovary-IN_GF_bovary	0.345881	FR_JV_balao-FR_JV_volta	0.500194
FR_GF_bovary-PT_JV_balao	0.855755	IN_JV_balao-PT_JV_centro	1.70813
FR_GF_bovary-FR_JV_balao	0.554391	IN_JV_balao-FR_JV_centro	0.677893
FR_GF_bovary-IN_JV_balao	0.826712	IN_JV_balao-IN_JF_centro	0.616145
FR_GF_bovary-PT_JV_centro	1.8716	IN_JV_balao-PT_JV_volta	0.396878
FR_GF_bovary-FR_JV_centro	0.564156	IN_JV_balao-IN_JV_volta	0.783818
FR_GF_bovary-IN_JF_centro	0.529365	IN_JV_balao-FR_JV_volta	0.320213
FR_GF_bovary-PT_JV_volta	1.1039	PT_JV_centro-FR_JV_centro	1.9029
FR_GF_bovary-IN_JV_volta	0.960825	PT_JV_centro-IN_JF_centro	1.5741
FR_GF_bovary-FR_JV_volta	0.970863	PT_JV_centro-PT_JV_volta	1.86471
IN_GF_bovary-PT_JV_balao	0.84781	PT_JV_centro-IN_JV_volta	1.68506
IN_GF_bovary-FR_JV_balao	0.533823	PT_JV_centro-FR_JV_volta	1.88119
IN_GF_bovary-IN_JV_balao	0.68434	FR_JV_centro-IN_JF_centro	0.403122
IN_GF_bovary-PT_JV_centro	1.92826	FR_JV_centro-PT_JV_volta	0.966821
IN_GF_bovary-FR_JV_centro	0.456607	FR_JV_centro-IN_JV_volta	0.530262
IN_GF_bovary-IN_JF_centro	0.46892	FR_JV_centro-FR_JV_volta	0.682566
IN_GF_bovary-PT_JV_volta	0.908663	IN_JF_centro-PT_JV_volta	0.887836
IN_GF_bovary-IN_JV_volta	0.873826	IN_JF_centro-IN_JV_volta	0.50894
IN_GF_bovary-FR_JV_volta	0.799248	IN_JF_centro-FR_JV_volta	0.756277
PT_JV_balao-FR_JV_balao	0.770515	PT_JV_volta-IN_JV_volta	1.0139
PT_JV_balao-IN_JV_balao	0.770727	PT_JV_volta-FR_JV_volta	0.567272
PT_JV_balao-PT_JV_centro	1.32566	IN_JV_volta-FR_JV_volta	0.775491

APÊNDICE H: Resultados do cálculo da distância entre textos para a classe IDIOMA

PARES DE TEXTOS *i* E *j*	$\delta_{i,j}$	**PARES DE TEXTOS *i* E *j***	$\delta_{i,j}$
ES_BG_marianela-ES_VI_barraca	0.579705	FR_AD_capitaine-IN_JA_love	0.932479
ES_BG_marianela-ES_JV_pasarse	0.459293	FR_AD_capitaine-IN_DD_robinson	0.250652
ES_BG_marianela-FR_AD_capitaine	0.838979	FR_AD_capitaine-PT_JA_viuvinha	1.62889
ES_BG_marianela-FR_GA_don	0.70377	FR_AD_capitaine-PT_MA_mao	0.980883
ES_BG_marianela-FR_PF_loup	0.59162	FR_AD_capitaine-PT_AA_cortico	1.17685
ES_BG_marianela-IN_LC_alice	0.885596	FR_GA_don-FR_PF_loup	0.914469
ES_BG_marianela-IN_JA_love	0.868149	FR_GA_don-IN_LC_alice	1.39336
ES_BG_marianela-IN_DD_robinson	0.85518	FR_GA_don-IN_JA_love	1.32725
ES_BG_marianela-PT_JA_viuvinha	1.19049	FR_GA_don-IN_DD_robinson	1.44107
ES_BG_marianela-PT_MA_mao	0.818322	FR_GA_don-PT_JA_viuvinha	1.28201
ES_BG_marianela-PT_AA_cortico	0.758782	FR_GA_don-PT_MA_mao	1.45767
ES_VI_barraca-ES_JV_pasarse	0.705186	FR_GA_don-PT_AA_cortico	0.52889
ES_VI_barraca-FR_AD_capitaine	1.06181	FR_PF_loup-IN_LC_alice	0.98387
ES_VI_barraca-FR_GA_don	0.707598	FR_PF_loup-IN_JA_love	0.892683
ES_VI_barraca-FR_PF_loup	0.76555	FR_PF_loup-IN_DD_robinson	0.738311
ES_VI_barraca-IN_LC_alice	1.09298	FR_PF_loup-PT_JA_viuvinha	1.31624
ES_VI_barraca-IN_JA_love	0.736048	FR_PF_loup-PT_MA_mao	0.942388
ES_VI_barraca-IN_DD_robinson	1.0817	FR_PF_loup-PT_AA_cortico	0.624574
ES_VI_barraca-PT_JA_viuvinha	1.10532	IN_LC_alice-IN_JA_love	0.726551
ES_VI_barraca-PT_MA_mao	1.14973	IN_LC_alice-IN_DD_robinson	0.878254
ES_VI_barraca-PT_AA_cortico	0.649126	IN_LC_alice-PT_JA_viuvinha	1.40473
ES_JV_pasarse-FR_AD_capitaine	0.649195	IN_LC_alice-PT_MA_mao	0.94433
ES_JV_pasarse-FR_GA_don	1.10518	IN_LC_alice-PT_AA_cortico	1.29487
ES_JV_pasarse-FR_PF_loup	0.648124	IN_JA_love-IN_DD_robinson	0.907749
ES_JV_pasarse-IN_LC_alice	0.846801	IN_JA_love-PT_JA_viuvinha	1.12324
ES_JV_pasarse-IN_JA_love	0.688035	IN_JA_love-PT_MA_mao	0.873774
ES_JV_pasarse-IN_DD_robinson	0.549039	IN_JA_love-PT_AA_cortico	1.10659
ES_JV_pasarse-PT_JA_viuvinha	1.37253	IN_DD_robinson-PT_JA_viuvinha	1.71282
ES_JV_pasarse-PT_MA_mao	0.600257	IN_DD_robinson-PT_MA_mao	0.880388
ES_JV_pasarse-PT_AA_cortico	1.03936	IN_DD_robinson-PT_AA_cortico	1.29353
FR_AD_capitaine-FR_GA_don	1.36365	PT_JA_viuvinha-PT_MA_mao	1.37582
FR_AD_capitaine-FR_PF_loup	0.631918	PT_JA_viuvinha-PT_AA_cortico	1.09584
FR_AD_capitaine-IN_LC_alice	0.860984	PT_MA_mao-PT_AA_cortico	1.34678