UNIVERSIDADE DA BEIRA INTERIOR
Engenharia

Caracterização Estética do Texto

(Versão Final Após Defesa)

Domingos Carlos Dionísio

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Doutor João Paulo da Costa Cordeiro
Co-orientador: Prof. Doutor Pedro Ricardo Morais Inácio

Covilhã, Novembro de 2018

Dedicatória

À minha querida esposa Wilsa Dionísio e às minhas dádivas Júlia Dionísio e Feliciano Dionísio, pelo amor dedicado e pelo apoio incomensurável.

Agradecimentos

Em primeiro lugar agradeço a Deus autor do meu destino, meu guia e socorro presente na hora da angústia, pelo Dom da Vida e por me ter permitido chegar a esta etapa da minha vida.
Agradeço aos Professores João Paulo Cordeiro e Pedro Inácio que, desde o início, acompanharam-me como orientadores científicos, inúmeras vezes souberam aturar as minhas fraquezas e, com muita paciência, mostraram-me os caminhos a seguir.
Ao Professor Dr. Mário Martins (da universidade Federal do Amapá), pela orientação e indicação de alguns artigos científicos relacionados com a minha dissertação.
Ao Dr. Mohsen Mesgar pelos valiosos esclarecimentos sobre o seu modelo de coerência do texto.
Aos meus colegas de turma e do Centro de Tecnologia da Linguagem Humana e Bioinformática, em especial ao Mário Pereira, Evaldo Chindele, Orlando Cawende, Domingos Oliveira, Edson Livongue, Fátima Dantas, Joaquim Mussandi, Ramos Pedro e Mvita Nzankulo, companheiros de longas noites, pela sua amizade, pelos bons momentos que partilhamos, pelas inúmeras trocas de experiência e pela forma como contribuíram direta ou indiretamente na redução do número de horas de trabalho que dediquei a esta dissertação.
Agradeço aos meus familiares e amigos mais próximos, em Especial ao meu pai Feliciano Dionísio, à minha mãe Júlia Cambundo e à minha esposa Wilsa Dionísio que, mesmo distante da pátria e de casa, confortaram-me e apoiara-me incondicionalmente.
Finalmente agradeço ao Instituto Superior Politécnico da Huila na pessoa do Professor Dr. Alberto Raimundo Wapota, pelo seu empenho e determinação que culminaram com a obtenção da Bolsa de Estudos financiada pelo INAGBE, sem esquecer a Universidade da Beira Interior por me abrir as portas e permitir que concluísse esta valiosa formação.

Resumo

Atualmente, a vasta quantidade de textos *online* à disposição de qualquer organização ou indivíduo tornou-se um enorme desafio. O consumidor/leitor encontra-se num labirinto de informações não estruturada (texto) em constante crescimento, muita dela de baixa qualidade. Em vários domínios, o leitor enfrenta um desafio ainda maior, sempre que necessitar de selecionar informações textuais confiáveis e de alta qualidade. É um processo trabalhoso, geralmente atingindo uma eficácia limitada. A avaliação sistemática e a seleção de documentos de alta qualidade tornaram-se uma tarefa impossível de ser executada manualmente por qualquer ser humano.

Portanto, o objetivo principal deste trabalho foi explorar os marcadores linguísticos que permitem uma caracterização efetiva da qualidade e estética contida no texto. Assim, apresentamos aqui uma análise exploratória e comparativa de um conjunto de vinte e um marcadores para avaliar a qualidade e a estética no texto. Também medimos o desempenho de auto-semelhança desses marcadores, em corpora, através de estimadores eficientes do parâmetro de Hurst. Quanto ao material experimental, reunimos e usamos dois corpora diferentes em termos de qualidade de escrita. Um corpus com alto padrão de qualidade, contendo obras clássicas da literatura, incluindo várias obras-primas do *Prêmio Nobel*. O outro com texto de qualidade inferior, vindo de vários *internautas*, incluindo textos escritos em *blogs* e por autores mais jovens e inexperientes. Os marcadores experimentados são agrupados em cinco tipos: vocabulário, coesão, coerência, complexidade sintática e figura de linguagem. As medições forneceram resultados muito interessantes, levando-nos a concluir que existem marcadores linguísticos de alta qualidade, muito úteis para distinguir entre textos bons e maus. O uso desses marcadores permitirá a implementação de sistemas capazes de realizar essa classificação, de forma automática e com alta qualidade.

Palavras-chave

Qualidade do Texto, Vocabulário, Coesão, Coerência, Complexidade Sintática, Auto-Semelhança e Figuras de Estilo.

Abstract

Nowadays the vast amount of online text at the disposal of any organization or individual has become a huge challenge. The consumer/reader finds itself in a maze of constantly growing unstructured information (text), much of it of poor quality. In several domains, the reader faces an even greater challenge, whenever he needs to select reliable and high-quality textual information. It is a laborious process, usually reaching a limited effectiveness. Systematically assessing and selecting high quality documents have become an impossible task to be performed manually by any human being.

Therefore, the main goal of this work was to explore the linguistic markers that enable an effective characterization of the quality and aesthetics contained in text. Thus, we present here an exploratory and comparative analysis of a set of twenty-one markers for assessing the quality and aesthetics in text. We have also measured the Self-Similarity performance of these markers, in corpora, through efficient estimators of the Hurst parameter. As for the experimental material, we have assembled and used two different corpora in terms of writing quality. One corpus having high quality standards, containing classic works of literature, including several Nobel-Prize masterpieces. The other one with text of inferior quality, coming from multiple internauts, including text written in blogs and by younger and inexperienced authors.

The experimented markers are grouped in five types: vocabulary, cohesion, coherence, syntactic complexity and figure of speech. Measurements provided very interesting results, leading us to conclude that there are high quality linguistic markers very useful for distinguishing between good and bad texts. The use of these markers will enable the implementation of systems capable of performing this classification, automatically and with high quality.

Keywords

Text Quality, Vocabulary, Cohesion, Coherence, Syntactic Complexity, Self-similarity, and Style Figures.

Índice

1 Introdução 1
1.1 Enquadramento . . . 1
1.2 Motivação . . . 2
1.3 Objetivos . . . 3
1.4 Abordagem e Contribuições . . . 3
1.5 Estrutura da Dissertação . . . 4

2 Estado da Arte 5
2.1 Qualidade e Estética do Texto . . . 5
2.2 Vocabulário na Qualidade e Estética do Texto . . . 6
2.2.1 Diversidade Lexical e o Comprimento do Texto . . . 6
2.3 Coesão e Coerência na Avaliação da Qualidade e Estética do Texto . . . 8
2.4 Complexidade Sintática na Avaliação da Qualidade e Estética do Texto . . . 11
2.5 Fractais na Avaliação da Qualidade e Estética do Texto . . . 12
2.6 Redes Complexas e Neurais Covolutivas na Avaliação da Qualidade e Estética do Texto . . . 12
2.7 Conjugação de Fatores na Avaliação da Qualidade e Estética do Texto . . . 13
2.8 Sumário . . . 14

3 Avaliação da Qualidade e Estética do Texto 15
3.1 Metodologia . . . 15
3.2 Marcadores do Texto . . . 17
3.2.1 Vocabulário . . . 17
3.2.2 Coesão e Coerência . . . 17
3.2.3 Complexidade Léxica e Sintática . . . 19
3.2.4 Recursos Estilísticos . . . 19
3.3 Métricas . . . 20
3.3.1 MTLD (**F1**) . . . 20
3.3.2 Taxa de Repetição (**F2**) . . . 22
3.3.3 Maturidade do Vocabulário (**F3**) . . . 23
3.3.4 Densidade Lexical (**F4**) . . . 24
3.3.5 Coerência de Entidades (**F5**) . . . 25
3.3.6 Transições de Entidades (**F6**, **F7**, **F8** e **F9**) . . . 29
3.3.7 Operadores Lógicos por Frase (**F10**) . . . 31
3.3.8 Densidade de Pronomes (**F11**) . . . 31
3.3.9 Sobreposição de Substantivos (**F12**) . . . 32
3.3.10 Média de Pronomes por Frase (**F13**) . . . 33
3.3.11 Similaridade entre Parágrafos Iniciais e Finais (**F14**) . . . 33
3.3.12 Similaridade entre Frases (**F15**) . . . 34
3.3.13 Sobreposição Média de Palavras (**F16**) . . . 34
3.3.14 Densidade de Pronomes por Substantivos (**F17**) . . . 35
3.3.15 Diversidade de Palavras de Conteúdo (**F18**) . . . 35
3.3.16 Comprimento Médio da Palavra (**F19**) . . . 36

3.3.17 Comprimento Médio da Frase (**F20**) . . . 36
3.3.18 Deteção de Anáfora (**F21**) . . . 36
3.4 Recursos Utilizados . . . 37
3.4.1 HultigLib . . . 37
3.4.2 OpenNLP . . . 38
3.4.3 Stanford CoreNLP . . . 38
3.4.4 Visual Interactive Syntax Learning . . . 39
3.4.5 New General Service List (NGSL) . . . 39
3.4.6 TestH . . . 40
3.4.7 Outros Recursos . . . 40
3.5 Sumário . . . 40

4 Experiências e Resultados 41
4.1 Conjunto de Dados . . . 41
4.1.1 Corpus de TExcel . . . 41
4.1.2 Corpus de TMeds . . . 43
4.2 Testes e Resultados . . . 43
4.2.1 Análise da Auto-Semelhança . . . 44
4.2.2 Análise Geral dos Textos . . . 45
4.3 Discussão dos Resultados . . . 46
4.3.1 Resultados da Primeira Experiência . . . 47
4.3.2 Resultados da Segunda Experiência . . . 47
4.4 Sumário . . . 50

5 Conclusão e Trabalho Futuro 51
5.1 Conclusão . . . 51
5.2 Trabalho Futuro . . . 52

Bibliografia 53

Lista de Figuras

3.1 Principais etapas do projeto. 15
3.2 *Bag of words.* 25
3.3 *Bag of words* de entidades. 25
3.4 Grafo de entidade ponderado. 26
3.5 Projeção unidirecional. 27
3.6 Grafo de entidade normalizado. 27
3.7 TC restrita para frases adjacentes. 28
3.8 Grafo de projeção normalizado. 28
3.9 Arquitetura geral da *Stanford CoreNLP*. Fonte: Manning et all [Manning2014]. . . 38
3.10 Texto resultado da anotação da frase ”O João foi buscar o seu carro ontem”pelo VISL. 39

4.1 Gráfico comparativo dos resultados obtidos na análise de fractalidade nos TExel e no TMeds. 45
4.2 Gráfico das médias obtidas pela aplicação das métricas dos marcadores do texto selecionadas. 46

Lista de Tabelas

3.1 Exemplo de matriz de incidência. 26

4.1 Distribuição do Corpora Utilizado. 41
4.2 Descrição detalhada das obras utilizadas na constituição do corpus de TExcel. . 42
4.3 Características da máquina utilizada para as experiências. 43
4.4 Distribuição dos textos por blocos de 20 frases utilizado nas medições de Auto-Semelhança. 44
4.5 Descrição dos valores do parâmetro de Hurst, por características do texto, obtidos pelas experiências realizadas na análise de fractalidade. 45
4.6 Descrição dos valores dos marcadores do texto selecionados, obtidos pelas experiências realizadas na análise geral dos corpora. 46
4.7 Descrição dos resultados do teste *t-student* para marcadores do vocabulário do texto. 47
4.8 Descrição dos resultados do teste *t-Student* para marcadores da coesão e coerência do texto. 48
4.9 Descrição dos resultados do teste *t-student* para marcadores da complexidade sintática. 49
4.10 Descrição dos resultados do teste *t-student* para marcadores de figuras de estilos. 50

Lista de Acrónimos

API	Application Programming Interface (Interface de programação de aplicações)
GPL	General Public License (Licença Pública Geral)
LD	Diversidade Lexical (*Lexical Diversity*)
HULTIG	Human Language Technology Information Group (Centro de Tecnologia da Linguagem Humana e Bioinformática)
Li-Fi	*Light Fidelity* (Fidelidade Luminosa)
LSA	*Latent semantic analysis* (Análise de Semántica Lantente)
MATTR	*Moving Average Type Token Ratio*
MLR	*Multiple Linear Regression* (Regressão Linear Múltiplo)
MTLD	Medida da Diversidade Lexical Textual (*Measure of Textual Lexical Diversity*
NGSL	Nova Lista Geral de Serviços (*New General Service List*)
OCR	*Optical Character Recognition* (Reconhecimento Ótico de Caracteres)
PLC	*Power Line Communication* (Comunicação de Linha de Energia)
PLN	Processamento da Linguagem Natural
TAACO	*The tool for the automatic analysis of text cohesion* (Ferramenta para a análise automática da coesão do texto)
TExcel	Textos excelentes
TMeds	Textos Médios e Mediocres
TTR	*Type-Token Ratio*
VISL	*Visual Interactive Syntax Learning* (Aprendizagem Visual Sintaxe Interativa)

Capítulo 1

Introdução

Neste capítulo faz-se uma apresentação do trabalho proposto, evidenciando os objetivos e as contribuições, bem como a importância do tema escolhido. Assim, o capitulo está estruturado da seguinte forma: **Secção 1.1** — Enquadramento; **Secção 1.2** — Motivação; **Secção 1.3** — Objetivos; **Secção 1.4** — Abordagem e Contribuições e **Secção 1.5** — Estrutura da Dissertação.

1.1 Enquadramento

Por meio do desenvolvimento de métodos modernos de transmissão de dados como a *Power Line Communication* (PLC) e a mais recente invenção - *Li-Fi*[1] verifica-se diariamente um aumento vertiginoso do volume de dados em formato de texto, produzido em um ritmo bastante acelerado. As páginas da *web*, *wikipedia*, *blogs*, portais de publicação de notícias, lojas *online*, redes sociais e digitação eletrónica de materiais textuais (como livros, revistas, jornais) por meio da tecnologia de Reconhecimento Ótico de Caracteres (OCR[2]), constituem exemplos práticos de como essa informação é produzida. Com o auxilio da *internet* todo texto produzido é disponibilizado (muitas vezes de forma livre) a qualquer indivíduo ou organização que dele necessitar. Por um lado, essa informação é bastante útil para muitos investigadores. Serve de matéria-prima em estudos científicos para diversas áreas, como por exemplo na Linguística Computacional, Psico-Linguística e Processamento da Linguagem Natural (PLN[3]). Por outro, a atenção especial a ter em conta é que nem todo texto é produzido por especialistas dotados de competência linguística. Muitas vezes é feito por indivíduos com idade reduzida, fraco nível académico e intelectual, entre outros fatores. Diante desse cenário torna-se inevitável colocar uma questão preponderante: há de facto qualidade em todo texto produzido e disponibilizado?
Convém aqui sublinhar que um texto não é apenas um emaranhado de palavras. Existem padrões estéticos na escrita que permitem distinguir se há ou não qualidade. Normalmente num bom texto pode-se notar se as frases estão de acordo com as regras gramaticais da língua, se há erros ortográficos e se existe uma articulação entre as frases, períodos e parágrafos. O ser humano facilmente consegue perceber esses padrões, bastando ser dotado de um nível linguístico e conhecimento do mundo suficiente para o efeito. Mesmo assim, nem sempre a avaliação do mesmo texto por dois ou mais especialistas tem o mesmo resultado, apesar de algumas vezes estar muito próximo.

[1] Acrónimo de "*Light Fidelity*"(traduzido literalmente por "Fidelidade da luz"). Refere-se a sistemas de comunicação com luz visível 5G que empregam luz *LEDs* para transmitir comunicações em alta velocidade, de forma similar como acontece no *Wi-Fi*. (https://pt.wikipedia.org/wiki/Li-Fi)

[2] Acrónimo de Optical Character Recognition, é uma tecnologia usada para o reconhecimento de caracteres provenientes de fontes diversas como imagens e documentos de vários formatos. O OCR faz o devido *escaneamento* e transforma o ficheiro original em um arquivo editável, o que se tornaria difícil com um *scaner* normal de documentos ou imagens. (https://www.abbyy.com/pt-br/ocr/)

[3] Subárea da Inteligência Artificial que estuda a capacidade e as limitações de uma máquina em entender a linguagem dos seres humanos. Tem como objetivo fornecer aos computadores a capacidade de entender e compor textos. (https://medium.com/botsbrasil/o-que-é-o-processamento-de-linguagem-natural-49ece9371cff)

Por esse pressuposto, algumas inquietações podem ser ainda levantadas: afinal o que está por trás da qualidade de um texto? Há atributos quantificáveis em textos que servem para aferir a sua qualidade?
Este trabalho aborda métodos para avaliar automaticamente a qualidade do texto. Propõe-se a experimentação de métricas capazes de detetar características inerentes aos textos e filtrar os atributos que garantem a sua qualidade.
Avaliar a qualidade do texto é bastante útil para variadíssimas situações. Na avaliação da escrita dos alunos de diferentes níveis de ensino, os professores têm o dever de analisar e avaliar a progressão dos mesmos e, por meio dos resultados obtidos, aplicar métodos próprios para superar as debilidades por eles apresentadas. Os professores devem fazer uso de ferramentas que permitam a realização desta tarefa de forma automática. Isso garante que os mesmos tenham mais tempo de ensinar e orientar os seus formandos, o que se torna bastante difícil quando é feito de forma manual.
Outro campo de aplicação é na avaliação de trabalhos científicos escritos por investigadores iniciantes. Os padrões de escrita de um bom artigo científico e de uma boa dissertação ou tese devem ser avaliados constantemente. Isto permite que o autor esteja ciente dos erros linguísticos que não devem ser cometidos e garante cada vez mais a produção de textos com a qualidade desejada.
No tocante a geração de texto através da sumarização automática[4], convém salientar que atualmente os textos obtidos ainda são de baixa qualidade. Os sistemas responsáveis pela sumarização de documentos devem ter a avaliação da qualidade do texto como uma etapa do processo e produzir resumos bons, no ponto de vista dos aspetos linguísticos, como por exemplo a gramática, ortografia, coesão e coerência. O mesmo princípio pode ser aplicado também em traduções de textos de uma língua para outra. Até ao presente momento, os *softwares* que permitem a realização desta tarefa apresentam insuficiência nesta vertente. Muitos deles fazem uma tradução literal, o que permite que as ideias do texto resultante não estejam suficientemente articuladas. Outro aspeto a ter em conta é a especificidade de cada língua. É importante que esses programas tenham métricas específicas de língua que permitem avaliar os textos traduzidos.
Os motores de busca constituem outro campo para aplicação das métricas de avaliação da qualidade e estética do texto. Para além da relevância dos termos de *input* na pesquisa, seria bom que os mesmos tivessem o módulo atinente a avaliação da qualidade. Apesar de existir muitas páginas retornadas numa consulta, nem todas têm a qualidade desejada. Um módulo com esse propósito permitiria que se obtivesse textos que atendam melhor as necessidades do pesquisador.

1.2 Motivação

Estudos sobre a avaliação do texto têm vindo a ser feitos há muito tempo. Existe uma variedade de métricas que permitem quantificar aspetos importantes inerentes a sua qualidade. Mas até ao momento, não existe um consenso sobre as propriedades que contribuem para tal.
Maior parte de trabalhos já realizados enquadram-se em dois cenários diferentes:

[4]É parte da aprendizagem de máquinas (*machine learning*) e da mineração de dados (*data mining*). Tem por objetivo encurtar um documento de texto (por meio de *softwares*), a fim de criar um resumo com os principais pontos do documento original. As tecnologias que podem fazer um resumo coerente levam em consideração variáveis como comprimento, estilo de escrita e sintaxe. (https://en.wikipedia.org/wiki/Automatic_summarization)

- Alguns fazem uma abordagem da qualidade do texto centrada no leitor. Essa abordagem gera bastante controversas. Por um lado, os leitores têm diferentes habilidades (como por exemplo nível académico, idade e conhecimento do mundo) e a perceção da qualidade de um texto fica bastante subjetiva. Por outro, existem aspetos atinentes à língua (e não ao leitor) que garantem a organização estética do texto. São os elementos gramaticais, de ortografia, da coesão e coerência do texto, entre outros. A principal deficiência desta abordagem é que as suas métricas podem ser bastante difíceis de serem automatizadas.
- Outras abordagens (mais ligadas a aspetos linguísticos) enquadram-se em uma das subáreas focadas na qualidade do texto: coesão e coerência, complexidade do texto, diversidade lexical e redes neuras e covolutivas. Trabalhos resultantes dessas correntes podem ser bastante limitados por estudarem atributos de apenas uma subárea.

Outro aspeto relevante é que maior parte das pesquisas sobre a avaliação da qualidade do texto são feitas em Inglês. Pouco ainda se fez na língua portuguesa. Por ser a sexta língua mais falada no mundo, existem muitos falantes do Português monolingue que podem fazer uso de sistemas que permitem avaliar textos escritos em Português.

1.3 Objetivos

O objetivo principal deste trabalho é experimentar modelos matemáticos que permitam avaliar qualitativamente Textos Excelentes (doravante TExcel) e Textos Médios ou Medíocres (doravante TMeds) e filtrar as marcações mais significativas que estão na base dessa distinção. Para a concretização do mesmo, foram definidos outros objetivos mais específicos, abaixo mencionados:

- Fazer uma revisão bibliográfica sobre o tema a que se propões, de forma a descrever os trabalhos mais relevantes sobre a avaliação da qualidade e estética do texto.
- Selecionar vários atributos do texto, com vista a serem aplicados nas experiências a realizar e filtrar aqueles que podem caracterizar TExcel (ver Secção 4.1.1) e TMeds (ver Secção 4.1.2).
- Construir uma aplicação em Java que permita concretizar as experiências pretendidas.
- Descobrir a possibilidade de existência de auto-semelhança no texto, através dos resultados obtidos pelas experiências a realizar.
- Escrever um artigo científico, fruto do trabalho e experiências propostas.
- Escrever a dissertação e apresentar as conclusões do trabalho.

1.4 Abordagem e Contribuições

Como já foi referido na Secção 1.2, maior parte das pesquisas já feitas têm enfoque na capacidade do leitor e em subáreas ligadas a qualidade do texto. As duas abordagens apresentam algumas insuficiências. Neste trabalho faz-se uma abordagem da avaliação do texto com base em elementos linguísticos presentes no texto. As métricas focadas no leitor ultrapassam o escopo dessa pesquisa. Desta feita, as contribuições deste trabalho assentam em alguns princípios que como:

1. **Automatizar métricas que capturam atributos focados em aspetos linguísticos, para aferir a sua qualidade e estética.** Existem características próprias do texto que podem ser quantificadas de forma automática. A sua análise permite filtrar aqueles que apresentam maior correlação com a qualidade do texto. Defende-se a ideia de que os resultados da aplicação das mesmas métricas em TExcel (ver Secção 4.1.1) e TMeds (ver Secção 4.1.2), garantem a obtenção desses atributos.

2. **Conjugação de fatores para uma melhor avaliação da qualidade e estética do texto.** A avaliação da qualidade do texto tem sido feita com base em aspetos de apenas uma sub-área. Uma avaliação mais abrangeste englobaria atributos do vocabulário, marcações do discurso (coesão e coerência), complexidade sintática, entre outros. Esta é a ideia principal desta pesquisa.

3. **Avaliar textos em Português através das métricas testadas na língua inglesa.** Estudos sobre a avaliação da qualidade do texto têm sido feito com maior profundidade na língua inglesa. Infelizmente, pouco se faz na língua portuguesa. Porém, um dos propósitos deste trabalho é experimentar métricas usadas no Inglês e descobrir se as mesmas podem der aplicadas nos textos em Português. Isso pode servir de base para futuras investigações em PLN, na avaliação da qualidade e estética do texto e nos aspetos linguísticos afetos ao Português.

4. **Análise e deteção de auto-semelhança.** Avaliar o texto e detetar a existência de estruturas fractais[5] pode ser bastante útil para a descoberta da sua beleza estética. O texto é normalmente composto por blocos de vários atributos e, estes, podem ser auto-semelhantes em diferentes partes (frases, períodos ou parágrafos). Este aspeto foi pouco explorado em pesquisas científica sobre a qualidade e estética do texto, o que faz deste trabalho um dos pioneiros nesta temática. O principal propósito é detetar se o grau de auto-semelhança das partes de um texto, estimado pelo parâmetro de Hurst, é proporcional a sua qualidade.

1.5 Estrutura da Dissertação

A dissertação está estruturada da seguinte forma:

- **no primeiro capitulo** foi introduzido o tema para que se perceba o enquadramento do mesmo, os objetivos e a motivação na base da sua escolha.
- **no segundo capitulo** descrevem-se os principais trabalhos relacionados com a avaliação da qualidade e estética do texto;
- **no terceiro capítulo** apresenta-se a proposta da avaliação da qualidade e estética do texto, que passa pela metodologia e métodos utilizados, descrição dos marcadores selecionados e dos algoritmos envolvidos;
- **no quarto capítulo** é feita a descrição do corpora utilizado, as experiências realizadas bem como a apresentação, analise e discussão dos resultados obtidos;
- **no quinto e último capítulo** apresentam-se as conclusões obtidas bem como as propostas de trabalho futuro.

[5] Um fractal é um objeto (e.g. geométrico) que pode ser subdividido em partes semelhantes ao original. Aqui, refere-se especificamente ao facto de que as propriedades estatísticas do texto são as mesmas, independentemente da escala através do qual é observado (Cordeiro etl [Cordeiro2015]).

Capítulo 2

Estado da Arte

O propósito deste capítulo é descrever os trabalhos mais relevantes sobre a qualidade e estética do texto. Existem diversas abordagens nas áreas de Linguística Computacional e de PLN, que visam a deteção de fatores que concorrem para avaliação da escrita. Tais abordagens utilizam métricas diversas para se atingir os objetivos preconizados. Infelizmente poucas pesquisas referenciam a conjugação destes fatores para avaliar a qualidade e estética do texto. Parte considerável das investigações têm enfoque em temas como: (i) Vocabulário; (ii) Coesão, (iii) Coerência; (iv) Complexidade Sintática; (v) Fractais; (vi) Legibilidade do Texto; e (vii) Redes Complexas e Neurais. No entanto, na sua maioria, são abordagens muito concentradas em poucas medidas.

A estrutura do capitulo tem a seguinte ordem: **Secção 2.1** — Qualidade e Estética do Texto; **Secção 2.2** — Vocabulário na Qualidade e Estética do Texto; **Secção 2.3** — Coesão e Coerência na Avaliação da Qualidade e Estética do Texto; **Secção 2.4** — Complexidade Sintática na Avaliação da Qualidade e Estética do Texto; **Secção 2.5** — Fractais na Avaliação da Qualidade e Estética do Texto; **Secção 2.6** — Redes Complexas e Neurais Covolutivas na Avaliação da Qualidade e Estética do Texto; **Secção 2.7** — Conjugação de Fatores na Avaliação da Qualidade e Estética do Texto; e **Secção 2.8** — Sumário.

2.1 Qualidade e Estética do Texto

Métodos automáticos para avaliar a qualidade da escrita de um texto podem ser bastante úteis para várias aplicações [Louis2012]. A grande questão que se coloca é quando considerar que um texto tem qualidade? Como já salientado, diversas investigações têm vindo a ser feitas com vista a demonstrarem quais fatores melhor concorrem para a distinção entre bons e maus textos. Em [Schriver1989], citado por [Pitler2008], faz-se uma abordagem da qualidade do texto com foco no leitor. Para a autora, a definição do que se pode considerar como um texto bem escrito e legível depende fortemente do nível linguístico do leitor. Essa tese é defendida também por vários investigadores que se debruçam sobre a legibilidade do texto ([Cavaco2010], [Chall1995], [Dubay2004]). Esta visão é bastante discutível pelo facto de que o mesmo texto pode ser avaliado como bom e mau, por dois leitores diferentes, dependendo do nível de conhecimento de cada um deles. Além disso, a competência linguística do leitor pode ser um fator insuficiente para avaliar a qualidade de um texto. Por exemplo, [Pitler2008] afirmam que um artigo científico muito bem escrito não será percebido como muito legível por uma pessoa leiga e um ótimo romance pode não ser apreciado por um aluno da terceira série[1].

Outro aspeto a ter em conta é que esta corrente limita em grande medida a tendência da avaliação do texto de forma automática. Na Linguística Computacional e em PLN, para avaliar

[1] No sistema educacional americano, a terceira série corresponde ao 3º ano de escolaridade do nível *elementar school* (primeiro nível), cuja primeira série os alunos frequentam-na aos seis anos de idade. Até a altura de ingressar ao nível superior os alunos frequentam doze séries. (https://studyusa.com/pt/a/28/o-sistema-de-educao-americano)

o texto são enfatizadas as propriedades que o mesmo apresenta e não as características do destinatário [Pitler2008].
Em [Cordeiro2015] considera-se um texto com qualidade, aquele que apresenta frases gramaticalmente corretas, alto grau lexical e estrutural e combinação retórica de palavras, frases e ideias. Este princípio é partilhado por [Oestling2017], segundo os quais o que faz um texto bom é uma confluência de qualidades diversas: narrativa coerente, gramática correta, ausência de erros ortográficos, um rico vocabulário e um conjunto de elementos linguísticos. Porém, [Louis2013] completa o conceito ao afirmar que a perceção de qualidade é influenciada por vários fatores: tópico interessante, conteúdo informativo, sem erros de gramática e ortografia, organização clara, escrita elegante e também bom *layout* e apresentação de texto na página.

2.2 Vocabulário na Qualidade e Estética do Texto

O vocabulário é um elemento fundamental no processo de escrita. Relativamente a um texto, a sua qualidade pode ser influenciada pelo nível de conhecimento lexical do escritor. Segundo Flower e Hayes [Flower1981], sem o vocabulário, as experiências e imagens sensoriais linguísticas que são armazenadas na memória a longo prazo, não podem ser comunicadas pelos escritores. Harmon et al [Harmon2005] citados em [Olinghouse2013], defendem que o vocabulário transmite conhecimento de conteúdo. Para compor um texto sobre determinado assunto, o escritor tem de conhecer o conjunto de palavras deste domínio e as funções sintáticas que as mesmas podem ter na frase. Na mesma linha, Laufer [Laufer2013] defende que não podemos nos comunicar em discurso oral ou escrito sem conhecer as palavras que transmitem as ideias que queremos expressar.
Apesar da importância incontestável que o vocabulário tem na composição, os seus marcadores têm sido pouco estudados na avaliação da qualidade e estética do texto. No seu trabalho sobre a densidade lexical e a diversidade lexical (em Inglês *Lexical Diversity* (LD)) na fala e na escrita, Johansson [Johansson2009] descobriu que as duas marcações fornecem informações importantes sobre os textos, no entanto, não podem ser usadas como a única maneira de julgar qualitativamente um texto. Natalie e Jacqueline [Olinghouse2009] usaram a diversidade de vocabulário (o mesmo que diversidade lexical), o vocabulário menos frequente, o comprimento médio de sílaba e o número de palavras polissilábicas como indicadores da qualidade nos textos narrativos. Os resultados indicaram que a diversidade lexical foi a mais estável e consistente das quatro variáveis de vocabulário. Olinghouse e Wilson [Olinghouse2013] estudaram a relação entre o vocabulário e a qualidade da escrita em três géneros (história, persuasivo e informativo). Eles usaram textos escritos por 105 alunos e marcaram as composições para a qualidade da escrita holística e várias construções de vocabulário diferentes: LD, maturidade, elaboração, palavras académicas, palavras de conteúdo e registo. Os resultados indicaram que os alunos variam o seu uso de vocabulário e as construções relacionadas à qualidade da escrita por género. Porém, a diversidade lexical é uma das marcações do vocabulário que tem sido mais explorado em diversos trabalhos relacionados.

2.2.1 Diversidade Lexical e o Comprimento do Texto

Segundo McCarthy e Jarvis [McCarthy2010], a LD refere-se ao conjunto de palavras diferentes usadas em um texto, com um maior alcance que indica uma maior diversidade. Muitas vezes

é referenciada como equivalente à riqueza lexical[2] [Daller2003]. É um indicador de desenvolvimento linguístico associado à quantificação da variação de palavras empregues num dado texto, ou seja, quanto maior a variação de palavras, maior a diversidade [Freitas2013]. Os autores afirmam que os índices de LD foram encontrados como indicativos de qualidade de escrita, conhecimento de vocabulário, competência falante, início de *Alzheimer*, variação auditiva e mesmo *status* socioeconómico.

O cálculo do índice LD foi inicialmente proposto por Templin [Templin1957a] como *type-token ratio* (TTR), ao examinar os resultados obtidos nas amostras de 50 enunciados escritos por 480 crianças entre 3 e 8 anos de escolaridade, fazendo dele a primeira métrica utilizada para a LD em linguagem escrita. O TTR consiste na razão entre o número de palavras diferentes (*types*) sobre o total de palavras (*tokens*) no texto.

$$TTR = \frac{types}{tokens} \quad (2.1)$$

Durante muito tempo, vários trabalhos com esta métrica foram feitos por diferentes investigadores como Richards ([Richards1987], [Richards1997]), Youmans [Youmans1990], Scherer et al [Scherer2002] e Thomas [Thomas2005]. Isso fez dela a medida considerada clássica para quantificar a LD.

Uma atenção particular dada ao comportamento da Equação 2.1 levantou questões quanto a eficácia da mesma. Para Martins [Martins2016], apesar da aparente eficácia desta medida em revelar a LD, existe uma relação de dependência com o comprimento do texto. O TTR demonstra sensibilidade às variações da dimensão do texto, isto é, quanto maior for o comprimento, menor será o valor do TTR. Este problema foi reportado em [Malvern2004] e em [McCarthy2010]. Isso levou muitos investigadores a subdividirem o texto em partes mais pequenas para a devida avaliação (por exemplo Biber [Biber1989]). Segundo McCarthy e Jarvis [McCarthy2010] , os trabalhos que não verificaram este pressuposto podem ter fornecido dados questionáveis como resultado (ver [Thomas2005], [Scherer2002] e [Ertmer2002]). As limitações do TTR levou ao o surgimento de várias abordagens visando contornar a situação. Como o TTR funciona bem para textos pequenos, uma proposta foi de se estabelecer um tamanho de amostra padronizado do texto a ser medido, mas nem isso serviu de solução porque (segundo Fergadiotis et al [Fergadiotis2013]) o grande problema é que, para que os resultados sejam comparáveis entre os estudos, os investigadores devem concordar com o numero de *tokens* necessários para medir o TTR. Pierre [Pierre1960] propôs em seus estudos a correção do TTR com o *Root* TTR (Equação 2.2), Herdan [Herdan1960] com o logaritmo do TTR (Equação 2.3) e Carroll [Carroll1964] com o TTR corrigido (Equação 2.4).

$$TTR = \frac{types}{\sqrt{tokens}} \quad (2.2)$$

$$TTR = \frac{\log types}{\log tokens} \quad (2.3)$$

[2]Riqueza lexical é um sinónimo de diversidade lexical, usado por vários investigadores (Por exemplo Daller et. al [Daller2003]) em PLN e não só. Mas por motivos de familiaridade e por ser mais usado na literatura, opta-se pelo uso da "diversidade lexical". Por outro lado, Malvern et al [Malvern2004] explicam que a medida LD é apenas uma parte da característica multidimensional da riqueza lexical.

$$TTR = \frac{types}{\sqrt{2tokens}} \quad (2.4)$$

Estas métricas foram criticadas por Durán et al [Duran2004], que demonstraram a ineficácia das mesmas na resolução do problema e consideraram-nas como meras transformações algébricas do TTR. No entanto, recentemente surgiram novas propostas, como os casos do Vocd-D (conhecido também por medida-D) proposto em [Richards1997], o *Moving Average Type Token Ratio* (MATTR) descrita em [Covington2010] e a Medida da Diversidade Lexical Textual (em Inglês *Measure of Textual Lexical Diversity*(MTLD)) descrito em [McCarthy2005].

Para Martins [Martins2016] a medida Vocd-D baseia-se num modelo de probabilidade para rastrear a diminuição da TTR conforme o número de *tokens* aumenta. Quanto maior for o valor D, maior será a diversidade lexical. Como ele, muitos autores utilizaram essa métrica nas suas pesquisas (por exemplo Berman e Verhoeven [Berman2002], McCarthy e Jarvis [McCarthy2007], Johansson [Johansson2009a], Malvern et al [Malvern2004] e Stromqvist et al. [Stroemqvist2002]). Contudo, McCarthy e Jarvis [McCarthy2010] testaram-no e detetaram dois inconvenientes: a) apresenta sensibilidade no comprimento do texto, o que contraria Richards e Malvern; b) o vocd-D apenas replica a função de distribuição hipergeométrica[3]. Através de avaliações de validade convergente (que consiste na avaliação de quão bom um índice concorda com outros índices que são amplamente reconhecidos como padrão), validade divergente (que consiste na avaliação de como um índice não concorda com índices que são considerados falíveis ou enganadores), validade interna (que consiste na avaliação da sensibilidade dos índices LD às variações no comprimento do texto, avaliadas por análises de correlação) e validade incremental (que consiste na avaliação do grau em que um dado índice é informativo acima e além de outro presumivelmente similar), McCarthy e Jarvis descobriram que esta medida comporta-se melhor em relação as demais e é única pouco afetada pelo problema do comprimento do texto [McCarthy2010]. Para medir a LD com vista a avaliação da qualidade de textos (devido ao tamanho variado dos mesmos), a melhor métrica a usar é o MTLD. Existem vários programas que permitem obter o índice da LD em textos (principalmente em Inglês). Porém para as métricas Vocd-D e MTLD é usado com maior frequência o *Textinspector*[4].

A métrica MTLD foi ainda pouco explorada por investigadores que trabalham com LD. Exemplos práticos (para além de McCarthy e Jarvis [McCarthy2010]) são os trabalhos de Treffers-Daller, J. [Treffers-Daller2013], Lieke Verheijen [Verheijen2016] e Koizumi e In'nami [Koizumi2012a]. Para o Português (até a elaboração do presente trabalho) não há nenhum registo com recurso a esta métrica.

2.3 Coesão e Coerência na Avaliação da Qualidade e Estética do Texto

Todo enunciado é sempre produzido com a intenção de se estabelecer a comunicação ou transmissão de informações para os leitores [Machado2012]. Partindo desse princípio, um texto não é apenas um aglomerado de palavras ou frases sem conexão. As mesmas precisam apresentar entre si uma relação que lhes confira sentido e também articulações gramaticais que dê clareza

[3]Ver Wu [Wu1993], para uma discussão da distribuição hipergeométrica

[4]Ferramenta *online* que permite fazer análise do texto e capturar dados relevantes para estudo do vocabulário. Entre várias métricas, o software usa o MTLD e o *Vocd-D*. Para mais informações consultar: https://textinspector.com/workflow (último acesso no dia 12/1/2018)

e precisão nas ideias apresentadas. A coesão e coerência são elementos essenciais de um texto com qualidade. Os seus níveis no texto afetam a compreensão dos leitores [Todirascu2013].
Graesser et al [Graesser2004] estabelecem uma distinção entre coesão e coerência. Segundo eles a coesão é uma característica do texto, enquanto a coerência é uma característica da representação mental do leitor em relação ao conteúdo do texto. A mesma ideia é defendida em [Crossley2016] ao afirmarem que, a coesão geralmente se refere à presença ou ausência de pistas explícitas no texto que permitem ao leitor fazer conexões entre as ideias expressas. Exemplos dessas sugestões explícitas incluem sobreposições de palavras e conceitos entre frases. Essas pistas comunicam ao leitor que as mesmas ideias estão sendo encaminhadas através de frases consecutivas. A coerência, em comparação com a coesão, refere-se à compreensão que o leitor retira do texto. É uma propriedade de textos bem escritos que os torna mais fáceis de ler e entender do que uma sequência de frases aleatoriamente encadeadas [Lapata2005]. Depende de uma série de fatores, incluindo pistas de coesão explícitas, sugestões de coesão implícitas (que estão mais intimamente ligadas à coerência do texto do que as pistas explícitas) e fatores não-linguísticos como conhecimento prévio e habilidades de leitura [Crossley2016]. A coesão está mais diretamente relacionada a aspetos que permitem conectar partes do discurso (e.g. frases, orações e parágrafos). A coerência se refere às articulações entre ideias no texto [Machado2012]; está vinculada ao texto, mas não depende somente dele [Freitas2013a]. Segundo Van Dijk [VanDijk1980], a coerência é uma propriedade semântica dos discursos, baseada na interpretação de cada frase individual em relação às outras frases.
Dijk [VanDijk1980], Charolles [Charroles1988], Dijk e Kintsch [Dijk1983] e Freitas [Freitas2013], estabelecem dois níveis de coerência, denominando-os de coerência local (que diz respeito às partes do texto como frases ou proposições e suas conexões lineares) e global (que se refere à totalidade do texto ou a fragmentos maiores, em que se estabelece uma conexão com o tema ou ideia central) [Freitas2013a]. O mesmo princípio é também aplicado à coesão do texto. Segundo Crossley et al [Crossley2016], para além da coesão local (verificada entre pequenos segmentos de texto, como a sobreposição nominal entre frases ou a vinculação de frases através de conectivos) e da coesão global (encontrada entre grandes partes de texto como sobreposição de substantivo entre parágrafos), existe também a coesão geral (que se refere à incidência de características de coesão em um texto inteiro).
A coesão e a coerência têm sido estudadas e aprofundadas por diversos investigadores a nível mundial e geralmente o principal objetivo consiste na avaliação da qualidade do texto analisado. O resultado de algumas investigações levou à criação de modelos e ferramentas de análise automática de textos, com intuito de facilitar os investigadores e utilizadores (e.g. professores e linguistas) nas suas investigações.
Graesser et al [Graesser2004] desenvolveram o Coh-Metrix[5]. Apesar de não ser totalmente grátis, é uma ferramenta de referência na literatura para processamento de texto em Inglês. Usa mais de 200 medidas de coesão, linguagem e legibilidade, por meio dos seus módulos que recorrem ao léxico, classificadores de *part-of-speech* (POS), analisadores sintáticos, *Latent Semantic Analysis* (LSA) e outros componentes que são amplamente usados na Linguística Computacional. A incidência de palavras de conteúdo (como substantivos, verbos, adjetivos e pronomes), de função (como pronomes, determinantes e preposições), a densidade de pronomes no texto (usada para medir a dificuldade de compreensão), proporção de operadores lógicos e os conectivos (de esclarecimento, aditivos, temporais e casuais), são algumas das medidas disponíveis nessa ferramenta. Vários investigadores fizeram uso dela nos seus trabalhos, como por exemplo

[5]Foi desenvolvido por investigadores da Universidade de Memphis. Para fins de uso pode-se aceder o endereço: http://www.cohmetrix.com/ (último acesso no dia 12/01/2018)

[McNamara2011], [Graesser2004] e [Toernqvist2015].

Todirasco et al [Todirascu2013] usaram 41 medidas para avaliar a coesão e a coerência de texto como indicadores de legibilidade. Eles usaram dois corpora, incluindo textos com diferentes níveis de dificuldade e mostram que algumas medidas coesivas (como pronomes pessoais, sintagmas nominais indefinidos, transição sujeito-objeto e objeto-objeto) são realmente bons indicadores de legibilidade do texto. As medidas usadas nessa investigação são relativas à POS (como a razão de pronomes por substantivos e proporção média de pronomes por frase), medidas de coerência lexical (como a média de similaridade, sobreposição de palavras e sobreposição de todos os lemas), coesão de entidade (como a transição sujeito-objeto e objeto—objeto), densidade de entidades (como média de entidades por frase e a proporção média de entidades únicas por documento) e cadeias de referência (como sintagmas nominais indefinidos, determinantes demonstrativos, pronomes demonstrativos e comprimento médio das cadeias de referência).

Crossley et al [Crossley2016] criaram uma ferramenta de analise de texto que recebeu o nome de *Tool for the Automatic Analysis of Text Cohesion*[6] (TAACO), de livre acesso e instalação no computador pessoal. A ferramenta incorpora mais 150 índices clássicos e outras mais recentes relacionados com a coesão do texto. As conclusões deste estudo fornecem a validação preditiva do TAACO e apoiam a avaliação de especialistas em coerência que afirmam não existir correlação entre a qualidade do texto e os índices de coesão local e geral, mas sim com os índices de coesão globais [Crossley2016]. O índice de pronomes, lemas de conteúdos repetidos, sobreposição de todos os lemas em cada duas frases, o TTR de lemas e a razão de substantivos por pronomes são algumas das medidas usadas e disponíveis na ferramenta.

Segundo Petler e Nenkova [Pitler2008] a maioria dos modelos que tentam capturar a coerência local entre frases foram baseadas ou inspiradas na teoria da centralidade (*centering*) [Grosz1995], que postularam fortes ligações entre o centro das atenções na compreensão de frases adjacentes, posição sintática e forma de referência. Barzilay e Lapata [Barzilay2008] fazem uma abordagem baseada no modelo Grade de Entidades. Esta sugere que o texto seja representado por uma matriz bidimensional (Grade de Entidades), que captura entidades do discurso nas frases. As linhas da grade correspondem às frases, enquanto as colunas às entidades do discurso. Baseando-se neste modelo, pesquisas relacionadas com coerência local para o Português (até à elaboração deste trabalho) podem ser encontradas em [Freitas2013], que aplicaram o modelo na implementação de um classificador para detetar quebras de linearidade que afetam a coerência de resumos científicos, [Silva2015] que também usaram o modelo (mas combinado com a estrutura retórica do discurso) para se detetar quebras de linearidade e mostrar mensagens específicas dos locais onde essas quebras são detetadas e [Freitas2013a] que investigou a aplicabilidade do modelo na avaliação de coerência de resumos científicos. Para o Inglês existe uma variedade de pesquisas que fazem uso do modelo ([Burstein2010], [Souza2013], [Cheung2010]). Uma particular atenção é dada aos trabalhos de Guinaudeau e Strube [Guinaudeau2013] e Mesgar e Strube [Mesgar2014], que fizeram uma adaptação ao modelo de Barzylay e Lapata [Barzilay2008] e criaram o modelo baseado em grafos e a sua versão normalizada respetivamente. Esse modelo supera a Grade de Entidades por resolver os problemas de dispersão de dados, dependência de domínio e complexidade computacional (encontrados no modelo de Barzilay e Lapata).

Para avaliação da coesão e da coerência do texto, existe uma variedade de ferramentas disponíveis. Entre elas destacam-se a *Intelligent EssayAssessor*[7] [Landauer2003], o *Criterion*[8] ([Higgins2004],

[6] Consultar: http://www.kristopherkyle.com/taaco.html (último acesso no dia 12/01/2018)

[7] https://images.pearsonassessments.com/images/assets/kt/download/IEA-FactSheet-20100401.pdf (último acesso no dia 19/01/2018)

[8] http://www.fairtest.org/facts/csrtests.html (último acesso no dia 20/01/2018)

[J.Burstein2003]) e o *Intellimetric*[9] [Elliot2003], que permitem avaliar a qualidade de textos escritos em Inglês. Para o Português infelizmente existem poucos recursos e sublinha-se aqui o *Scientific Portuguese*[10] e o *SciPo* ([Freitas2013], [Souza2013]). Entre os vários recursos disponíveis, o *SciPo* possui um módulo de análise que deteta potenciais problemas de coerência em resumos científicos [Freitas2013].

2.4 Complexidade Sintática na Avaliação da Qualidade e Estética do Texto

Segundo Crossley e McNamara [Crossley2014], a complexidade sintática refere-se à sofisticação e a variedade das formas sintáticas produzidas por um orador ou escritor (em consonância com [Lu2011] e [Ortega2003]). Estudos sobre a sua correlação com a qualidade do texto tem gerado controvérsias entre investigadores durante décadas. O resultado negativo de estudos sobre a relação entre sintaxe e qualidade de escrita permitiu uma fraca atenção pelos investigadores e consequentemente a redução considerável em pesquisas relacionadas. Apesar desta redução muitos trabalhos dedicaram-se em avaliar esta correlação.

As descobertas de Hunt [Hunt1970] determinaram a existência de três medidas de complexidade sintática - cláusulas por unidade-T (originalmente *T-unit*[11]), palavras por cláusula e palavras por unidade-T - que eram vistas como indicadores mais confiáveis de aumento da maturidade na escrita. A unidade-T é considerada como o método tradicional de medir a complexidade sintática e foi inicialmente usada para avaliar o desenvolvimento da escrita em escritores nativos da língua (L1) [Crossley2014]. Outras medidas de complexidade lexical podem ser encontrados em ferramentas computacionais, como o *MAT*[12] e *Coh-Metrix*.

Segundo Crowhurst [Crowhurst1983] (citado em [Beers2009]) o comprimento da unidade-T e da cláusula não são suficientemente discriminatórios para servir como indicadores confiáveis de qualidade de escrita. Por outro lado, Stewart e Grobe [Stewart1979] encontraram uma relação fraca entre o comprimento da frase e a qualidade do texto. Beers e Nagy [Beers2009] contradizem a conclusão de Crowhurst afirmando que o próprio estudo de Crowhurst [Crowhurst1983] sobre o relacionamento entre a complexidade sintática e a qualidade da escrita encontrou resultados mistos, mas as suas descobertas deixam aberta a possibilidade de que a relação entre a complexidade e a qualidade sintática dependa de idade e género. No seu estudo, que visou explorar a relação das três medidas de complexidade sintática propostas por Hunt para a qualidade do texto, Beers e Nagy usaram textos escritos por 43 estudantes na sétima ou oitava série (23 meninos e 20 meninas). Os seus resultados indicaram que a complexidade sintática está relacionada à qualidade do texto para escritores adolescentes, mas essa relação depende do género de texto e da medida específica da complexidade sintática utilizada.

Uma pesquisa mais recente sobre a relação entre a complexidade sintática e a qualidade do texto foi protagonizada por Thilagha Jagaiah [Jagaiah2017]. Foi usada a Análise Factoral de

[9] http://www.vantagelearning.com/products/intellimetric/intellimetric-how-it-works/ (último acesso no dia 20/01/2018)

[10] http://www.nilc.icmc.usp.br/scipo/classif.php?componente=resumo (último acesso no dia 21/01/2018)

[11] Na linguística , o termo unidade T foi cunhado por Kellogg Hunt. É definido como as "frases gramaticalmente permitidas mais curtas nas quais a escrita pode ser dividida, ou uma unidade minimamente terminável". Muitas vezes, mas nem sempre, uma unidade T é uma frase. fonte: https://en.wikipedia.org/wiki/T-unit

[12] https://sites.google.com/site/multidimensionaltagger/home (Último acesso no dia 30/01/2018)

Confirmação (originalmente *Confirmatory Factor Analysis*[13] (CFA)) para testar um modelo de hipóteses de 28 medidas de complexidade sintática e quatro variáveis latentes: padrão de frase, comprimento de frase, conector de frase e sofisticação de frase. No total foram usados dados de 1.029 testes argumentativos anotados com a ferramenta *Coh-Metrix* (versão 3.0). Para examinar a relação, Jagaiah aplicou um modelo de Regressão Linear Múltiplo (originalmente *Multiple Linear Regression*[14] (MLR)) e os resultados indicaram uma modesta relação positiva entre cada uma das quatro variáveis latentes e a qualidade de escrita.

2.5 Fractais na Avaliação da Qualidade e Estética do Texto

Uma área pouco aprofundada no processamento automático de texto é a de fractais. De facto, não há nenhum registo em Português sobre a correlação deles com a qualidade do texto e pouco ainda se fez na língua inglesa, o que torna uma área a ser explorada por investigadores. Uma característica importante dos fractais e de carácter interessante para estudos em PLN é a auto-semelhança. Segundo Cordeiro et al [Cordeiro2015], a auto-semelhança refere-se à propriedade de um processo estocástico parecer estatisticamente idêntico para qualquer escala (agregação) a partir da qual é observado. No seu estudo sobre beleza fractal no texto, os autores procuraram descobrir se vários atributos que caracterizam partes do texto (e.g. frases) são auto-semelhantes. Os autores usaram cinco corpora escritos em Inglês de três géneros diferentes – Literatura, História e noticias de *Blogs*. Para análise dos dados aplicaram os testes de normalidade de Kolmogorov-Smirnov e medição dos parâmetros de Hurst. Os resultados indicam a existência de uma beleza fratal no texto produzido por humanos (gerada de forma inconsciente) e sugerem que a qualidade do texto é diretamente proporcional ao grau de auto-semelhança, contudo ainda é necessária uma análise mais complexa e exaustiva. Até à elaboração desta pesquisa, não se observou evidências de outro trabalho semelhante.

2.6 Redes Complexas e Neurais Covolutivas na Avaliação da Qualidade e Estética do Texto

As redes complexas são provenientes da mecânica estatística e representam uma extensão da teoria dos grafos. Têm uma grande importância na área de computação, pelo facto da sua utilização exigir pouca ou nenhuma engenharia de recursos manuais [Oestling2017]. Por esta e outras particularidades, este conceito tem sido usado em aplicações de PLN, na avaliação da qualidade de texto ([Antiqueira2005], [Antiqueira2007], [Oestling2017]).
O texto pode ser visto como uma rede, onde cada *type* representa um nó e as palavras adjacentes as arestas [Antiqueira2005]. Como as redes incorporam as associações mais imediatas entre palavras e conceitos, a sua topologia — quantificada por várias medidas, como o grau do nó, o coeficiente de agrupamento e o caminho mais curto — pode fornecer informações sobre algumas propriedades do texto, como estilo e autoria [Oestling2017]. Em [Antiqueira2005], o texto foi modelado como uma rede complexa com o objetivo de codificar as relações entre os conceitos de um texto. Neste modelo, para cada par de palavras consecutivas (nós), existe uma respetiva aresta direcionada na rede, com um peso que corresponde ao número de vezes em que há associações destas no texto. Depois dos testes efetuados chegaram à conclusão que

[13] https://en.wikipedia.org/wiki/Confirmatory_factor_analysis (Último acesso no dia 28/01/2018)
[14] https://www.wsj.com/europe (Último acesso no dia 20/12/2078)

os parâmetros das redes complexas apresentam uma evidente correlação com a qualidade de textos e, portanto, são potencialmente úteis para distinguir textos bons e maus. As descobertas de Antiqueira et al [Antiqueira2005] foram reforçadas em [Antiqueira2007] em que se avaliou os textos produzidos por alunos do ensino médio em Português, com base em três fatores de qualidade: coesão e coerência (ver Secção 2.3), adequação às convenções de escrita e o desenvolvimento do tema. Os textos foram representados como redes livres de escala (modelo de adjacência de palavras), dos quais foram obtidas características de rede típicas, como o grau de entrada/saída, o coeficiente de agrupamento e o caminho mais curto. Segundo os resultados obtidos dos testes, a coesão e a coerência mostraram a correlação mais forte com a qualidade do texto, em relação aos outros dois critérios. Os autores consideram que provavelmente esse resultado indica que as medições de rede são capazes de capturar como o texto é desenvolvido em termos dos conceitos representados pelos nós nas redes [Antiqueira2007].
Num trabalho mais recente [Oestling2017] os autores fazem uma abordagem da avaliação da qualidade do texto com a aplicação de redes neurais convolutivas. Este modelo, de aprendizagem automática, utiliza um corpus de texto gerado pelo escritor (de qualidade variável) e um corpus de texto contrastante (considerado de alta qualidade). Os testes foram feitos com o uso de um grande corpus de ensaios de estudantes universitários (anotado manualmente) e um corpus longitudinal da produção escrita de alunos de linguística. Os resultados indicam que o modelo é capaz de fornecer avaliações de qualidade locais em diferentes partes de um texto, o que permite um *feedback* visual sobre a localização das partes problemáticas, bem como a maneira de avaliar quais recursos de texto são capturados.

2.7 Conjugação de Fatores na Avaliação da Qualidade e Estética do Texto

Poucos trabalhos fazem referência à conjugação de fatores para avaliar a qualidade de textos. O primeiro estudo que teve em consideração o uso de vários recursos linguísticos na predição da qualidade do texto foi de Pitler e Nenkova [Pitler2008]. Eles usaram características léxicas, sintáticas e discursivas, para produzir um modelo altamente preditivo de julgamentos dos leitores humanos da legibilidade do texto e descobriram que as relações do discurso estão fortemente associadas à qualidade percebida do texto. Para o seu estudo, os autores usaram textos do *Wall Street Journal*[15] destinados a uma audiência adulta educada, para analisar os fatores de legibilidade, incluindo o vocabulário, a sintaxe, a coesão, a coerência das entidades e o discurso. Pitler e Nenkova chegaram a conclusão que o discurso e o vocabulário são os fatores mais fortemente ligados à qualidade do texto.
Num estudo mais recente [Louis2012], sublinha-se que tal como em muitas pesquisas, os métodos utilizados anteriormente processam diferenças entre bons e maus textos como percebidos por um único nível de audiência (e.g. leitores adultos ou com certo nível de escolaridade). Para contrapor essa ideia, propôs uma abordagem que considera um único público, concentração em quatro aspetos ligados a qualidade do texto (o conteúdo/tópico discutido, a gramática da frase, a coerência do discurso e o estilo de escrita) e as particularidades do género do texto estudado (publicações académicas, artigos de notícias sobre ciência e textos gerados por máquinas, em particular o resultado dos sistemas automáticos de resumo de texto). Segundo a autora, esses aspetos tornam a investigação da qualidade do texto linguísticamente interessante porque, por

[15] https://www.wsj.com/europe (Último acesso no dia 20/12/2078)

definição, o foco está em uma ampla gama de propriedades do texto em vez de adequação para um leitor (Louis [Louis2012]). Na mesma linha, McNamara et al [McNamara2010] usaram o *Coh-Metrix* para examinar o grau em que os testes de alta e baixa proficiência podem ser preditos pelos índices linguísticos de coesão (i.e., correferência e conectivos), complexidade sintática (por exemplo, número de palavras antes do verbo principal, estrutura de frases sobrepostas), a diversidade de palavras usadas pelo escritor e as características das palavras (por exemplo, frequência, clareza, imaginabilidade). Nesse estudo os autores descobriram três bons factores indicadores de qualidade do texto (complexidade sintática a diversidade lexical e a frequência de palavras) e os resultados indicam que as características textuais que caracterizam a boa escrita não estão alinhadas com os recursos que facilitam a compreensão de leitura. São as características linguísticas associadas a dificuldades do texto e a sofisticação da linguagem.

2.8 Sumário

Descreveram-se neste capítulo as principais abordagens usadas na avaliação da qualidade/complexidade de um texto. Como se pode constatar, maior parte dos trabalhos já realizados, particularizam a avaliação em atributos de uma subárea apenas. Porém, é fundamental que a avaliação da qualidade de texto seja mais abrangente possível, com vista a torna-la mais eficiente.
No Capítulo 3 apresenta-se o método de avaliação da qualidade e estética do texto proposto. A proposta enquadra-se na última abordagem (Secção 2.7), visto ser a menos explorada pelos investigadores e a que melhores garantias confere na avaliação da qualidade de escrita. Para além do Inglês, são analisados também textos escritos em Português. Propõe-se aqui experimentar e verificar até que ponto as métricas testadas em Inglês podem também ser aplicadas na língua portuguesa.

Capítulo 3

Avaliação da Qualidade e Estética do Texto

Nesse capítulo faz-se uma descrição pormenorizada sobre a ideia central do presente trabalho, a metodologia seguida desde a delimitação do tema até as conclusões, bem como os métodos e procedimentos envolvidos na obtenção dos dados por meio de experiências. Assim, o capítulo está estruturado da seguinte forma: **Secção 3.1** — Metodologia; **Secção 3.2** — Marcadores do Texto; **Secção 3.3** — Métricas; **Secção 3.4** — Recursos Utilizados e **Secção 3.5** — Sumário.

3.1 Metodologia

O objetivo desse trabalho é de experimentar modelos que visam avaliar a qualidade e estética do textos e filtrar os marcadores mais significativos que estão na base da distinção dos TMeds e TExcel. Foram selecionados aspetos referentes às propriedades do texto e não a atributos extra linguísticos, que também são importantes nesta tarefa. A sequência de passos que descreve as principais etapas envolvidas neste processo é representada na Figura 3.1.

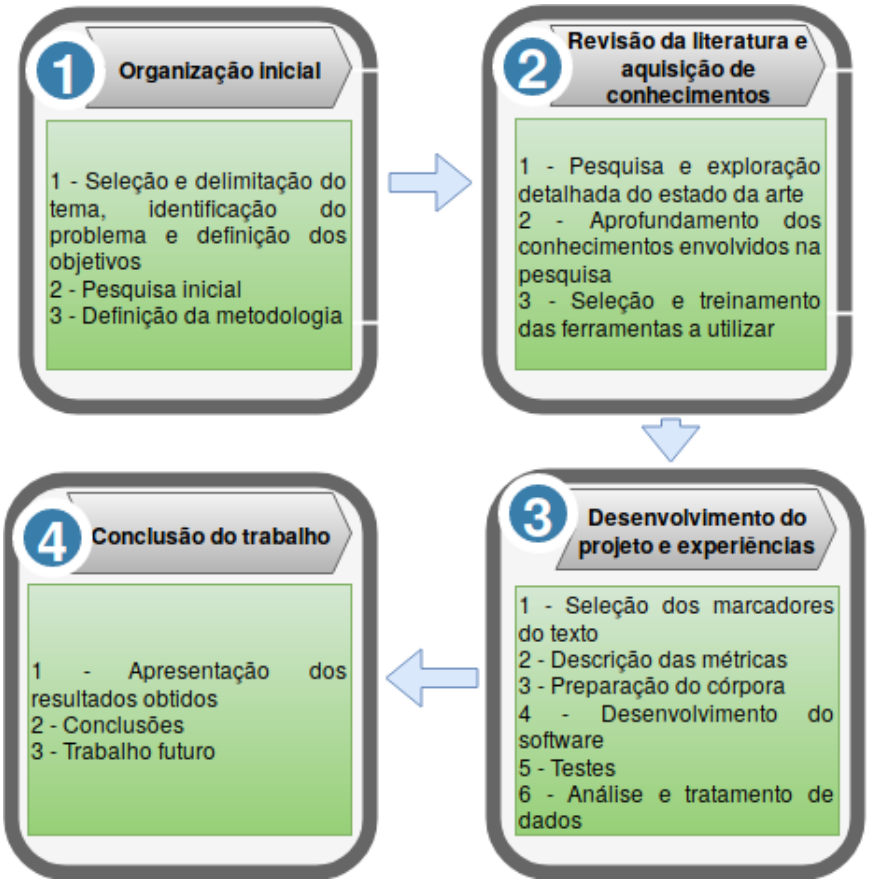

Figura 3.1: Principais etapas do projeto.

- **Primeira etapa:** Marcou o início da pesquisa, com o propósito de permitir a escolha do tema

e determinar os primeiros passos. Como se pode constatar na Figura 3.1, as atividades principais feitas nessa fase foram: i) seleção e delimitação do tema; ii) identificação do problema e definição dos objetivos; iii) pesquisa inicial sobre o tema proposto; iv) definição da metodologia adotada no trabalho. Esta etapa permitiu a familiarização com o tema proposto e constituiu a base a partir da qual dependeram as fases seguintes.

- **Segunda etapa:** Permitiu aprofundar os conhecimentos sobre o tema e as principais áreas envolventes, bem como a escolha da ferramenta e das bibliotecas a serem utilizadas. As principais atividades feitas foram: i) pesquisa e elaboração do estado da arte; ii) aprofundamento dos conhecimentos envolvidos na pesquisa, isso é, domínio das subárea relacionadas com o tema; iii) escolha da ferramenta de desenvolvimento e treinamento sucessivos na mesma. Esta etapa é de particular importância por ser quase permanente ao longo do projeto. Nela foi possível efetuar os ajustes no projeto até a obtenção dos resultados apresentados na última etapa. O método usado nesta etapa foi a pesquisa bibliográfica e teve como base, a revisão de trabalhos de pendor científico já publicados, tais como, revistas científicas, artigos, livros monografias, dissertações e teses.

- **Terceira etapa:** Constituiu a principal fase do presente trabalho. Nela foi proporcionada toda a parte prática por meio do método experimental (para obtenção dos dados) e do método estatístico (para análise e tratamento dos dados). Como ilustrado na Figura 3.1, as principais atividades desenvolvidas foram: i) seleção dos marcadores to texto, bem como as respetivas métricas para a sua quantificação; ii) constituição dos corpora, com base no qual foram feitos os testes; iii) cdesenvolvimento da aplicação na plataforma escolhida; iv) testes; v) análise e tratamento dos dados.

- **Quarta etapa:** Marcou o fim do projeto, cujas principais atividades foram: i) apresentação dos resultados; ii) apresentação das conclusões; e iii) descrição do trabalho a ser realizado no futuro.

Coletou-se vários textos em diversas fontes e, com eles, formaram-se dois corpora: um para TExcel (ver Secção 4.1.1) e outro para TMeds (ver Secção 4.1.2). Considerou-se TExcel aqueles pertencentes a escritores notabilizados, que já foram congratulados com o *Prémio Nobel da Literatura* e do *Prémio Oceanos*. Os TMeds, são aqueles escritos por vários *internautas* e por adolescentes com pouca experiência na escrita. A descrição detalhada pode ser lida na Secção 4.1.

Para o processamento dos corpora e aplicação das medidas foi desenvolvida uma pequena aplicação em Java e utilizaram-se as bibliotecas *HultigLib*[1] (ver Secção 3.4.1), *OpenNLP*[2] (ver Secção 3.4.2) e *Stanford CoreNLP*[3] (ver Secção 3.4.3). Uma das dificuldades enfrentadas em todo o processo foi a análise sintática para textos em Português. Não foi encontrada nenhuma biblioteca em Java com funções que permitissem programar e executar tal tarefa na aplicação construida. Esta situação foi contornada com a anotação do texto através do *Visual Interactive Syntax Learning*[4] (ver Secção 3.4.4). Ao texto resultante extraíram-se os elementos necessário (por exemplo o sujeito e objeto) e aplicou-se nas medidas correspondentes.

[1] Biblioteca de PLN da Universidade da Beira Interior. Para mais informações consultar em `http://www.di.ubi.pt/$\sim$jpaulo/hultiglib/` e `https://github.com/johnycordeiro/hultiglib`

[2] https://opennlp.apache.org/

[3] https://nlp.stanford.edu/software/lex-parser.html

[4] Projeto de pesquisa e desenvolvimento no Instituto de Linguagem e Comunicação (ISK) da Universidade do Sul da Dinamarca (SDU). Para mais informações consultar o endereço https://visl.sdu.dk/

Para algumas medidas utilizaram-se os dicionários em Português e em Inglês do projeto *OpenOffice*[5], que permitiram filtrar algumas palavras desconhecidas (principalmente nos TMeds), desde que não fossem entidades nomeadas. A fim de detetar palavras entre as menos comuns na língua, recorreu-se ao *New General Service List* [6] para o Inglês (ver Secção 3.4.5) e a lista de palavras comuns do Português disponibilizada pelo *quizlet.com*[7] para o Português.

3.2 Marcadores do Texto

Um dos objetivos fundamentais de um escritor em relação a um texto por ele produzido é estabelecer comunicação com o leitor. O texto é visto como um conjunto de várias palavras articuladas e de sinais de pontuação que, respeitando as regras gramaticais da língua usada, pode servir de meio para exteriorizar o pensamento do escritor sobre determinado assunto. Olhando por esse ângulo seria fácil avaliar a qualidade do texto quando o leitor e o escritor têm o mesmo nível de conhecimento (linguístico e do mundo). Infelizmente, na maioria dos casos, este pressuposto não se verifica. Isso acontece por estarem envolvidos vários fatores que garantem que o texto não seja visto apenas como um amontoado de palavras, o que torna bastante difícil a sua avaliação qualitativa.
Para o presente trabalho foram selecionados 21 marcadores que serviram de base para as experiências propostas. Muitos deles já foram bastante utilizados em estudos de PLN (ver Secção 2) e podem servir de indicadores bastante valiosos para a avaliação da qualidade de escrita. Para simplificação do texto, esses marcadores serão descritos por **F1**, **F2**, ... e **F21** respetivamente. Conforme mencionado na Secção 1.4 a abordagem proposta é baseada na conjugação de vários fatores para julgar a qualidade e estética do texto. Pensamos que esse método garante uma maior abrangência de fatores linguísticos a serem usados e, consequentemente, uma avaliação mais eficiente e segura. Com base na literatura e nos trabalhos anteriores, as 21 variáveis selecionadas podem ser agrupadas em quatro categorias principais:

3.2.1 Vocabulário

Os marcadores do vocabulário, embora alguns estudos os tenham ignorado, constituem elementos importantes na avaliação de escrita. O vocabulário pode traduzir o nível lexical do escritor numa determinada língua. Para a sua avaliação foram selecionadas a LD (**F1**), medida pela MTLD proposta em [McCarthy2005], com o propósito de se determinar a variação de palavras utilizadas; a taxa de repetição (**F2**) para obter o nível de repetição de palavras; a sofisticação do vocabulário (**F3**) para se medir o nível de maturidade apresentado, isto é, a utilização de palavras de pouca frequência na língua corrente [Olinghouse2013] e a densidade (**F4**) que visou obter a proporção de palavra de conteúdo no texto [Johansson2009].

3.2.2 Coesão e Coerência

Como descrito na Secção 2.3, a Coesão e a Coerência são dois elementos bastante importantes para a compreensão de um texto. Ajudam no entendimento de como as ideias se encontram articuladas.

[5] https://extensions.openoffice.org/en
[6] http://www.newgeneralservicelist.org/
[7] https://quizlet.com/140548978/1000-palavras-mais-usadas-em-portugues-em-geral-flash-cards/

A Coesão garante a ligação entre os constituintes de um texto, com base na gramática, para tornar compreensível as ideias que o escritor pretende transmitir. É um recurso linguístico que consiste no uso adequado de articulações gramaticais como conjunções, alguns advérbios e preposições, para conexão harmoniosa entre as palavras, frases, períodos e parágrafos de um determinado texto. É normalmente usada como uma das propriedades de Coerência Textual.
A Coerência é um recurso linguístico com um certo grau de complexidade. Diferente da Coesão, não depende somente dos elementos textuais. Como afirma Machado [Machado2012], a Coerência se refere às articulações entre ideias presentes no texto.
No presente trabalho adotaram-se quatro subcategorias que correspondem aos níveis de Coerência e de Coesão local, global e geral, descritas na Secção 2.3.

1. **Coerência Baseada em Entidades**
   Para capturar os níveis de Coerência do texto por meio de entidades nomeadas (**F5**), usou-se o modelo baseado em grafos normalizados [Mesgar2014] e selecionamos a projeção P_{Acc}[8]. Através dela medimos a coerência entre frases adjacentes por meio dos papeis sintáticos (sujeito e objeto) das entidade nas frases. Seguindo o modelo Grade de Entidades de Barzilay e Lapata [Barzilay2008] e do modelo baseado em grafo de Guinaudeau e Strube [Guinaudeau2013], foi possível obter as frequências de transições das entidades entre as frases no texto, isto é, as funções sintáticas das entidades nas frases n e n + 1: sujeito (S) ou objeto (O). O propósito foi descobrir no texto a probabilidade das transições S $\rightarrow$ S (**F6**), S $\rightarrow$ O (**F7**), O $\rightarrow$ O (**F8**) e O $\rightarrow$ S (**F9**), em consonância ao trabalho de Todirascu et all [Todirascu2013].

   As transições S $\rightarrow$ S permitem detetar a probabilidade de uma entidade com papel de sujeito na frase n aparecer na frase n + 1 também como sujeito. A transição S $\rightarrow$ O mede a probabilidade de uma entidade com papel de sujeito na frase n existir na frase n+1 como objeto. A transição O $\rightarrow$ O determina a probabilidade de uma entidade que é objeto na frase n poder ter um papel sintático de objeto na frase n+1. A última transição (O $\rightarrow$ S) permite obter a probabilidade de uma entidade ser objeto na frase n e sujeito na frase n+1.

   Neste trabalho foram consideradas de entidades nomeadas todos os substantivos (nomes próprios), tais como nomes de pessoas, locais, animais, organizações, entre outros, que se encontram mencionados no texto. Para deteta-los é necessário segmentar o texto em frases e cada uma delas em *Part-of-speech (POS) tagging*, para categorizar cada *token* na sua classe gramatical (substantivo, pronome, verbo, entre outros). A deteção desses elementos no texto, foi feita com os recursos encontrados nas bibliotecas utilizadas (ver Secção 3.4), nas versões correspondentes. Devido às pesquisas que têm vindo a ser feitas nessa área, os resultados podem ser melhores quanto mais atual for a versão da biblioteca.

2. **Coesão Lexical**
   Permite a reiteração dos mesmos constituintes do texto (e.g. palavras) nas frases, que resulta numa adequação semântica ao tema desenvolvido. A Coesão lexical pode ocorrer por mecanismos de repetição ou por substituição. A Coesão por repetição consiste em repetir o mesmo elemento textual (e.g. uma palavra) em diferentes frases, formando uma

[8]No modelo baseado em grafo de Guinaudeau e Strube [Guinaudeau2013] essa projeção permite capturar as transições de entidades em frases adjacentes, através dos papeis sintáticos (sujeito e objeto) que as mesmas têm. Esses papeis têm os pesos três (quando a entidade é um sujeito), dois (quando a entidade é um objeto) e um (quando a entidade não possui nenhum dos papeis anteriores). P_{Acc} é a expressão utilizada para se referir a este tipo de projeção.

harmonia no texto. A Coesão por substituição tem o mesmo efeito em relação à anterior, a diferença é que, em vez de se utilizar os mesmos elementos, a reiteração é feita através de mecanismos de sinonímia/antonímia, hiperonímia/hiponímia e holonímia/melonímia.

Algumas medidas dessa categoria foram experimentadas no trabalho de Todirascu et all [Todirascu2013] e os seus resultados forneceram evidências de uma possível correlação com a qualidade de escrita. Neste trabalho destacaram-se a sobreposição de substantivos (**F12**) (também usada por Crosley et all [Crossley2016]), para determinar a reiteração do mesmo substantivo em várias frases subsequentes; a semelhança semântica entre frases iniciais e finais (**F14**), que permitiu obter a similaridade lexical entre frases iniciais e finais; similaridade média entre frases adjacentes (**F15**) (também usada por Graesser et all [Graesser2004]), utilizada para quantificar a média da similaridade lexical entre par de frases adjacentes; e a média de sobreposição de palavras entre frases (**F16**) (também usada por Crosley et all [Crossley2016]), para detetar o número médio de palavras entre duas frases consecutivas.

3. **Baseadas em POS**
   As partes do discurso são elementos importantíssimos para a Coesão e a Coerência do texto. Neles se sobressaem os pronomes que permitem, entre muitos aspetos, encandear as ideias e a consequente progressão do texto. Neste trabalho foram selecionados a densidade de pronomes (**F11**) ([Graesser2004] e [Crossley2014]), para quantificar a incidência de pronomes em relação as outras classes gramaticais; a proporção média de pronomes por frases (**F13**) ([Pitler2008] e [Todirascu2013]), para se determinar a média de pronomes em cada frase; a proporção de pronomes por substantivos (**F17**) ([Pitler2008] e [Todirascu2013]) , para se achar a razão entre pronomes por substantivos em cada frase; e a diversidade de palavras de conteúdo (**F18**) [Crossley2014], para quantificar a variação da distribuição de palavras de contudo ao longo do texto.

3.2.3 Complexidade Léxica e Sintática

As medidas de complexidade léxica e sintática têm sido usadas com poucos trabalhos na avaliação de escrita, permite avaliar a composição sintática das frases no texto. Para esta categoria foram selecionadas apenas três medidas para avaliar o seu impacto na qualidade e estética do texto. A primeira é a proporção média de operadores lógicos por frase (**F10**) (também usada por Graesser [Graesser2004]), para determinar a proporção média de operadores lógicos por frases. O comprimento médio da frase (F19) e a da palavra (**F20**) (ambos utilizados por Pitler [Pitler2008]), para determinar as médias de palavras por cada frase e de carateres por cada palavra respetivamente.

3.2.4 Recursos Estilísticos

As figuras de estilos (ou recursos estilísticos) podem influenciar na qualidade e estética do texto. Normalmente são usadas para dar mais expressividade às frases. Estudos mais aprofundados sobre esta temática serão feitos num futuro trabalho. Para esta dissertação foi selecionado apenas uma das figuras de estilo — anáfora (**F21**) — por influenciarem nos níveis de repetição de palavras ou expressões no texto.

3.3 Métricas

3.3.1 MTLD (**F1**)

Como já referido na Secção 2.2, a MTLD [McCarthy2010] permite obter a diversidade do vocabulário do texto (ver Algoritmo 1). Cada *token* do texto é usado para calcular o TTR (ver Equação 2.1) até se atingir o fator predefinido, e.g. 0.72.

A avaliação da LD (ver Secção 2.2) com MLTD é sequencial, o que significa que cada palavra do texto é usada para o cálculo do TTR. Numa primeira fase é estabelecido o valor do fator. Seguidamente calcula-se o TTR (usando a Equação 2.1) até que o seu valor alcance o fator. Neste momento adiciona-se 1 ao contador (que inicialmente tem valor 0) e reinicia-se o TTR (ver instruções de 5 à 18 do Algoritmo 1). O processo é repetido de forma sequencial até ao comprimento do texto (e.g. com a última palavra). Como o último bloco do texto pode não ser suficiente para se formar um fator, se assim acontecer, o seu TTR é usado para se calcular o valor percentual em relação ao intervalo $[0.72, 1.00]$ (ver instruções de 14 à 17 do algoritmo 1). Posteriormente é adicionado ao contador para se formar o valor total dos n fatores. Este valor é usado como denominador na Equação 3.1 para obter primeiro valor do MTLD . O procedimento é feito de forma normal (da primeira à última palavra) e reversa (da última à primeira palavra) e o valor final do MTLD é a média dos dois valores (MTLD normal e reverso), conforme a a equação 3.2.

Algorithm 1 Pseudo código para o calculo da MTLD

```
1:  Inicio
2:  function CALCULARMTLDIR(tipo)
3:      listaW ← palavras do texto
4:      n ← numWords
5:      for w ∈ listaW do
6:          listaWords ← listaWords ∪ {w}
7:          listaTypes ← listaTypes ∪ {w}
8:          TTR ← totalTypes/totalWords
9:          if TTR ≤ 0.72 then
10:             fatores ← fatores + 1
11:             listaWords ← ∅
12:             listaTypes ← ∅
13:         end if
14:         if w = ultimaPalavra ∧ TTR ≥ 0.72 then
15:             fatorParcial ← (1.0 − TTR)/(1.0 − 0.72)
16:             fatores ← fatores + fatorParcial
17:         end if
18:     end for
19:     mtldIR ← n/fatores
20:     Return mtldIR
21: end function
22: function CALCULARMTLD()
23:     mtldInverso ← CALCULARMTLDIR(normal)
24:     mtldReverso ← CALCULARMTLDIR(reverso)
25:     mtld ←(mtldNormal+ mtldReverso) / 2
26:     Return mtld
27: end function
28: Fim
```

$$MTLD_{normal,reverso} = \frac{tokens}{\sum_{i=1}^{N} TTR_i + \frac{1-TTR_{final}}{1-0.72}} \quad (3.1)$$

$$F1 = \frac{(MTLD_{normal} + MTLD_{reverso})}{2} \quad (3.2)$$

Tomemos como exemplo os seguintes enunciados:

- **"O que as vitórias têm de mau é que não são definitivas. O que as derrotas têm de bom é que também não são definitivas."** (José Saramago)

- **"Sempre chega a hora em que descobrimos que sabíamos muito mais do que antes julgávamos."** (José Saramago)

- **"Um dia você aprende que as verdadeiras amizades continuam a crescer, mesmo a longas distâncias. E o que importa não é o que você tem na vida, mas quem tem na vida. Aprende que não temos que mudar de amigos, se compreendermos que os amigos mudam."** (Verónica Shoffstall)

No primeiro exemplo obtém-se:

Para o MTLD normal, o número de fatores é 1, obtido do primeiro TTR = 0,71. Ou seja:

TTR = O (1/1) que (2/2) as (3/3) vitórias (4/4) têm (5/5) de (6/6) mau (7/7) é (8/8) que (8/9) não (9/10) são (10/11) definitivas (11/12). O (11/13) que (11/14) as (11/15) derrotas (12/16) têm (13/17) de (14/18) bom (15/19) é (15/20) que (15/21)

Último TTR = 1, isto é, "também (1/1) não (2/2) são (3/3) definitivas (4/4)"
Como 1 > 0.72, calcula-se a percentagem correspondente no intervalo [0.72 - 1]:
(1-1)/(1-0,72) = 0
O MTLD normal é então calculado como:
47 / (1 + 0) = 47

Para o MTLD iverso, o número de fatores também é 1, obtido do primeiro TTR = 0,70. Ou seja:
TTR = definitivas (1/1) são (2/2) não (3/3) também (4/4) que (5)5 é (6/6) bom (7/7) de (8/8) têm (9/9) derrotas (10/10) as (11/11) que (11/12) o (12/13) definitivas (12/14) são (12/15) não (12/16) que (12/17)
Último TTR = 1, isto é, "é (1/1) mau (2/2) de (3/3) têm (4/4) vitórias (5/5) as (6/6) que (7/7) o (8/8)"
Como 1 > 0.72, (1-1)/(1-0,72) = 0.
O MTLD reverso é então calculado como:
47 / (1 + 0) = 47

Logo, MTLD = (47 + 47) / 2 = 47

No segundo exemplo o MTLD normal é dado por: 15 / (0 + 0,5) = 30. Por sua vez o MTLD reverso também é dado por: 15/ (0 + 0,5) = 30
Logo, MTLD = (30 + 30) / 2= 30
No último exemplo o MTLD normal é dado por: 47 / (1 + 0,85) = 25,4. Por sua vez o MTLD reverso é dado por: 47/ (1 + 0,64) = 28.65
Logo, MTLD = (25,4 + 28.65) / 2= 27.02

Ao contraio de outras medidas da LD, o MTLD não tem um valor limite. A interpretação que dele se faz é que quanto maior for, mais diversificado é o vocabulário do texto.

3.3.2 Taxa de Repetição (**F2**)

Foi utilizada para a simples computação do nível repetição do vocabulário (Algoritmo 2).

$$F2 = \frac{tokens - types}{tokens} - anaforas \quad (3.3)$$

Com anáforas < $\frac{tokens-types}{tokens}$. O cálculo de anáforas é feito mediante a Equação 3.26 e a sua computação baseou-se no algoritmo 14.
Par a aplicação dessa medida é necessário um pré-processamento do texto, que consiste na remoção das *stop words* e dos sinais de pontuação. As *stop words* são comummente conhecidas por palavras de função e pelo facto de serem usadas com maior frequência nos textos, podem constituir fator tendencioso para esta métrica.

Algorithm 2 Pseudo código para a taxa de repetição

1: **Inicio**
2: $tokens \leftarrow$ palavras do texto sem stop words
3: $repeticoes \leftarrow$ totalTokens - types
4: $taxaDeRepeticao \leftarrow$ (repeticoes / totalTokens) - nivel de anaforas
5: **Return** taxaDeRepeticao
6: **Fim**

Para diminuir o peso da redundância no vocabulário, foi subtraído do quociente entre os *tokens* e *types*, o valor do nível de anáforas presente no texto (**F21**), conforme ilustrado na Equação 3.3.
Como exemplo tem-se a frase de Manuel Bandeira:

"Terra da castanha, terra da borracha, terra de beribá, bacuri, sapoti"

Depois da remoção das *stop words* tem-se:

"terra castanha terra borracha terra beribá bacuri sapoti"

A taxa de repetição será dada por: [(8 - 6) / 8] - 0.25 = 0. Onde 0.25 corresponde quantificação de anáforas presente no texto.
Isso significa dizer que, apesar de haver palavras repetidas, o nível de repetição é nulo por causa da figura de linguagem empregue.

3.3.3 Maturidade do Vocabulário (**F3**)

Esta métrica é uma adaptação inspirada no trabalho de Olighouse e Wilson [Olinghouse2013] e permite obter o nível de proficiência do vocabulário (ver Algoritmo 3). Baseou-se em capturar palavras desconhecidas na língua (***wd***), palavras comuns (***wc***), palavras elaboradas (***we***) e quantificar os seus níveis no texto. Originalmente, Olighouse e Wilson buscam apenas capturar no texto o vocabulário menos frequente, que é o quociente entre as palavras que não estão entre as mais comuns e o vocabulário.
Convém salientar que detetar palavras elaboradas pode ser uma tarefa bastante complexa, porque, alguns autores, podem incorporar gírias ou calão nos seus textos. Palavras dessas duas categorias, podem ser confundidas por ***we*** e resultar em valores altos para essa métrica. O uso de dicionários para verificar a existência de palavras na língua foi o recurso usado para contornar essa situação.
O cálculo de **F3** baseou-se na Equação 3.4. Como se pode notar, o valor de ***we*** foi usado como penalização.

$$F3 = \frac{\sum_{i=1}^{N} wc_i}{N} + \frac{\sum_{j=1}^{N} we_j}{types} - \frac{\sqrt{\sum_{x=1}^{N} wd_x}}{N} \quad (3.4)$$

Onde, ***wc*** são as palavras comuns, ***wd*** são as palavras desconhecidas, ***we*** são as palavras mais elaboradas e **N** o total de palavras no texto.
As palavras comuns para o Inglês foram obtidas da *New Words Service List* e para o Português, devido a inexistência de uma lista semelhante, utilizou-se uma lista de palavras usadas pelo *quizlet.com* no ensino da língua portuguesa em aprendizes e falantes não nativos. Para a deteção das **we** utilizaram-se dicionários do projeto *OpenOffice* para verificar a existência de cada uma delas, sempre que não pertencesse em nenhum dos outros dois grupos (**wc** e **wd**).

Algorithm 3 Pseudo código para a maturidade do vocabulário

Inicio
$listaW \leftarrow$ lista de palavras do texto
$N \leftarrow |listaW|$
$types \leftarrow$ vocabulario do texto
$lwc \leftarrow$ lista de Wc
for $w \in$ listaW do
    if w $\in$ lwc then
        $wc \leftarrow wc + 1$
    else if w $\in$ DicOpenOffice then
        $we \leftarrow we + 1$
    else
        $wd \leftarrow wd + 1$
    end if
end for
$F3 \leftarrow (wc/N) + (we/types) - sqrt(wd)$ /N
Return F3
Fim

Como exemplo tem-se as seguintes frases:

"Os vivos são pó levantado, os mortos são pó caído; os vivos pó que anda, os mortos pó que jaz" (Padre Vieira)

"Passa mbora o meu mambo, vou guerar wi"

A primeira frase é totalmente composta de palavras conhecidas, o que permite obter um bom valor para a maturidade do vocabulário usado: ((10/20) + (0/10) - $\sqrt{0}$/20) = 0,5. Já o segundo exemplo, contém algumas palavras estranhas e que são desconhecidas na língua portuguesa. Na verdade a expressão correta seria: **"Dá-me o meu objeto, vou-me embora amigo"**. Devido as palavras "estranhas"(mbora, mambo, guerar e wi), resultaria num valor relativamente baixo para o **F3**, isto é, (4 /8) + (0/ 4) - ($\sqrt{4}$/8) = 0,25.
A interpretação que se faz do resultado da **F3** é que quanto mais próximo de 1 ou superior, melhor é a proficiência do vocabulário e mais próximo de 0, pior é.

3.3.4 Densidade Lexical (**F4**)

Esta métrica permite descrever a proporção de palavras de conteúdo (substantivos, verbos, adjetivos e muitas vezes advérbios) em detrimento das palavras de função (ver Algoritmo 4). É de fácil computação e consiste na razão entre o total de palavras de conteúdo pelo comprimento do texto [Johansson2009]. No nosso trabalho, em vez de utilizar o número de *tokens*, optou-se em trabalhar com *types*. O motivo é que o texto pode conter muita repetição de palavras e isso pode tornar o resultado tendencioso quanto ao comprimento do texto. Assim, a aplicação da medida baseou-se na a equação 3.5.

$$F4 = \frac{1}{n}\sum_{i=1}^{n} wct_i \quad (3.5)$$

Onde, wct é cada um dos *types* de conteúdo e n o vocabulário.

Algorithm 4 Pseudo código para a densidade lexical

```
Inicio
types ← vocabulario do texto
for w ∈ types do
    if w ∈ Palavra de Conteudo then
        wct ← wct + 1
    end if
end for
F4 ← wct/n
Return F4
Fim
```

Exemplo:

"Tudo o que um sonho precisa para ser realizado é alguém que acredite que ele possa ser realizado." (Roberto Shinyashiki)

Das 18 palavras do exemplo obtém-se um vocabulário de 14 delas. Depois do processo de POS obtém-se 8 palavras de conteúdo (sonho, alguém, precisa, acredite, ser, realizado, possa e é). Logo, F4 corresponde a 0,57, ou seja, $\frac{8}{14}$. O resultado é um valor no intervalo [0, 1] e indica que quando mais próximo de 1, o texto possui uma alta densidade de palavras de conteúdo e mais próximo de 0, a densidade é baixa.

3.3.5 Coerência de Entidades (**F5**)

A Coerência local no texto pode ser obtida através dos pesos das entidades entre frases adjacentes (ver Algoritmo 5). Como já referido na Secção 3.2.2, para medir como as frases se encontram articuladas por meio de entidades, foi usado o modelo baseado em grafo normalizado [Mesgar2014] devido algumas vantagens que apresenta e os seus resultados serem relativamente melhores aos da *Grade de Entidades* [Barzilay2008].
Este modelo representa o texto como um grafo bipartido ou de entidades G= (V_s, V_e, L, W) da matriz de incidência (tabela 3.1). Para tal, é necessário segmentar o textos em frases (*sentence segmentation*) e, posteriormente, *tokenizar* cada uma delas, formando um *bag of words* por cada frase do texto.

Como exemplo tem-se o seguinte enunciado:

"A espantosa realidade das coisas é a minha descoberta de todos os dias. Cada coisa é o que é. E é difícil explicar a alguém quanto isso me alegra, e quanto isso me basta. Basta existir para se ser completo." (Fernando Pessoas)

A *sentence segmetation* e a *tokenization* originariam os seguintes *bag of word* por frase:

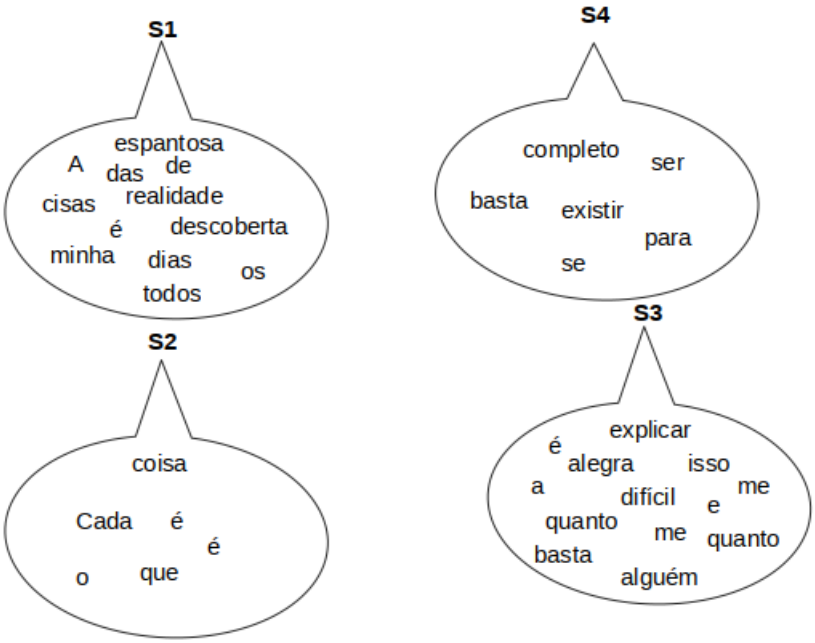

Figura 3.2: *Bag of words.*

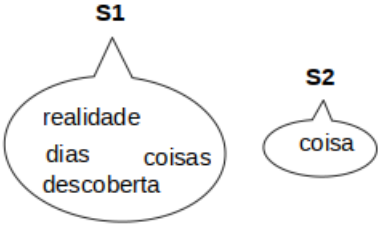

Figura 3.3: *Bag of words* de entidades.

Na matriz de incidência (ver Tabela 3.1), as linhas representam as frases do texto e as colunas as o vocabulário de entidades.

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}
S_1	1	2	1	0	0	0	0	0	0	0
S_2	0	3	1	1	2	0	0	0	0	0
S_3	0	0	0	3	0	1	0	1	0	0
S_4	0	0	0	0	0	0	0	3	1	0
S_5	0	0	0	0	0	0	0	0	1	2

Tabela 3.1: Exemplo de matriz de incidência.

Esta matriz fornece informações relevantes para se construir o grafo de entidades ponderado, como se pode ver na Figura 3.4.

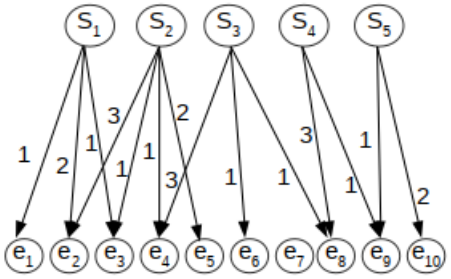

Figura 3.4: Grafo de entidade ponderado.

O grafo bipartido é definido por um conjunto de nós V_s que representam as frases e outro conjunto V_e que representa as entidades. Os nós são ligados pelas arestas L associadas a pesos W. Se numa célula C_{ij} da matriz de incidência tiver um peso $W(e_j, S_i)$, é criada uma aresta entre a frase S_i e a entidade e_j no grafo bipartido. O valor do peso na aresta depende do papel sintático de e_j em S_i. Se e_j for mencionada mais de uma vez em S_1 escolhe-se o papel sintático com mais relevância (e.g. sujeito $>$ objeto $>$ outro).

$$W(e_j, S_i) = \begin{cases} 3 \; se \; e_j \; for \; sujeito \; na \; frase \; S_i \\ 2 \; se \; e_j \; for \; objeto \; na \; frase \; S_i \\ 1 \; se \; e_j \; for \; diferente \; de \; sujeito \; e \; objeto \; na \; frase \; S_i \end{cases}$$

A projeção unidirecional P_{Acc} captura as relações existentes entre as frases no texto. O peso da aresta entre uma frase S_i e outra adjacente S_{i+1} é dado por:

$$W_{ii+1} = \sum_{e \in E} w(e, S_i) * w(e, S_{i+1}) \quad (3.6)$$

Onde, $w(e, S_*)$ é o peso da entidade e na frase S_* e E é o conjunto de entidades das frases (S_i e S_{i+1}).

O resultado é um grafo de projeção unidirecional do grafo de entidade, com as frases do texto ligadas por arestas associadas aos pesos correspondentes, como se pode ver na Figura 3.5.

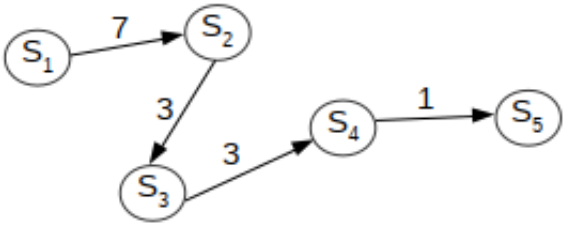

Figura 3.5: Projeção unidirecional.

A normalização dos pesos é feita através da divisão de cada um deles pelo grau do nó (ver Figura 3.6). Este procedimento incorpora a informação negativa (i.e. entidades que não ocorrem em outras frases) e captura a importância relativa das entidades e das frases respetivamente. A importância da entidade e na frase S_i é então calculada com base a Equação 3.7:

$$Imp(e, S_i) = \frac{w(e, S_i)}{\sum_{e \in E} w(e, S_i)} \quad (3.7)$$

Onde, $w(e, S_i)$ é o peso da entidade e na frase S_i e E é o conjunto de entidades na frase S_i. Este procedimento permite obter o grafo de entidades com pesos normalizados, conforme é ilustrado na Figura 3.6.

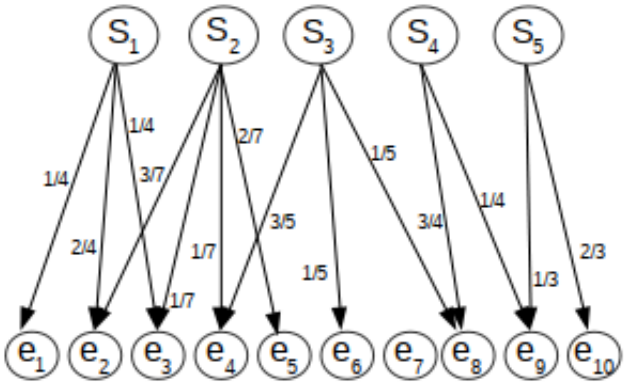

Figura 3.6: Grafo de entidade normalizado.

A normalização da projeção unidirecional é feito com a incorporação de um nó virtual em frases adjacentes (ver Figura 3.7). As funções de pontuação $lw_{S_i}^{S_{i+1}}$ e $lw_{S_{i+1}}^{S_i}$ correspondem aos pesos locais da frase S_i de acordo com frase S_{i+1} (e vice-versa) e são calculados com base as Equações 3.8 e 3.9. O valor obtido é igual ao grau de cada frase dividido pela soma dos graus das duas frases adjacentes. A Equação 3.10 permite calcular a projeção unidirecional normalizada para cada par de frases do texto.

$$lw_{S_i}^{S_{i+1}} = \frac{te(S_i)}{te(S_i) + te(S_{i+1})} \quad (3.8)$$

$$lw_{S_{i+1}}^{S_i} = \frac{te(S_{i+1})}{te(S_{i+1}) + te(S_i)} \quad (3.9)$$

Onde, $te(S_*)$ é o número de entidades na frase S_*.

$$W_{S_{i_{i+1}}} = \sum_{e \in E} \left\{ [lw_{S_i}^{S_{i+1}} * Imp(e, S_i)] + [lw_{S_{i+1}}^{S_i} * Imp(e, S_{i+1})] \right\} \quad (3.10)$$

Para impedir que o modelo seja influenciado pelo tamanho das frases, é feita a aproximação da saliência da entidade e na frase S_i pelo produto do grau $lw_{S_i}^{S_{i+1}}$ de S_i com a importância $Imp(e, S_i)$ de e em S_i.

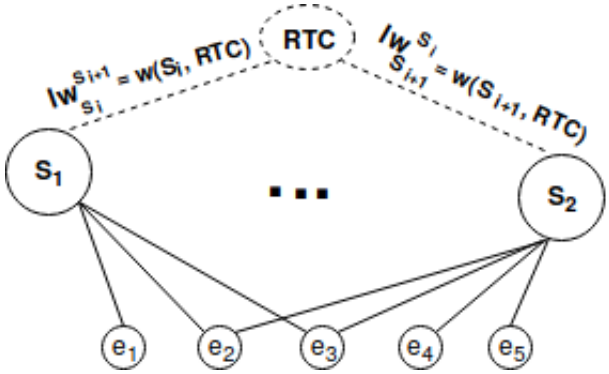

Figura 3.7: TC restrita para frases adjacentes.

O peso de uma aresta entre frases adjacentes S_i e S_{i+1} é a soma dos pesos das arestas da cada frase com o no virtual restrito, com base na Equação 3.10. Este procedimento resulta em projeções com pesos normalizados (ver Figura 3.8), utilizados para se obter a coerência local em todo texto.

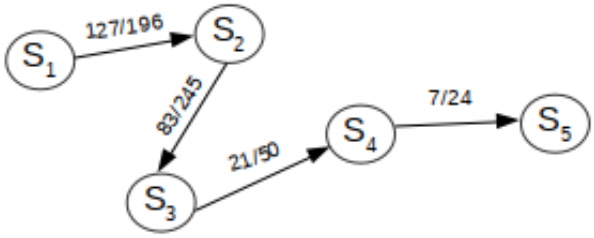

Figura 3.8: Grafo de projeção normalizado.

Este grafo fornece informações necessária para se calcular a coerência local baseada em entidades, através da Equação 3.11.

$$F5 = \frac{1}{N} \sum_{i=1}^{N} OutDegree(S_i) \quad (3.11)$$

Na Equação 3.11, $OutDegree(S_i)$ é a soma dos pesos das arestas no grafo de projeção normalizado e N é o número de frases no texto. Esse valor também pode ser visto como a soma dos valores da matriz de adjacência do grafo de projeção normalizado, dividido pelo número de frases do texto.
Por exemplo, os valores do grafo de projeção normalizado (ver Figura 3.8) originam uma coerência local de 0,33968. Sabendo que o resultado de F5 é um valor no intervalo [0, 1,00], o número obtido fornece evidências de que o texto tem uma razoável coerência local baseada em entidades do discurso.

Para esta métrica, o texto é dito coerente quanto mais próximo de 1 for o valor obtido e, incoerente, quanto mais próximo de 0.

Algorithm 5 Pseudo código para coerência local

Inicio
function PESODAENTIDADE($frase$, $entidade$)
    if $entidade \in$ sujeito then
        $peso \leftarrow 3$
    else if $entidade \in$ objeto then
        $peso \leftarrow 2$
    else
        $peso \leftarrow 1$
    end if
    return $peso$
end function
function SOMATORIODEPESOS($frase$)
    $somatorio \leftarrow 0$
    for $e_i \in$ frase do
        $somatorio \leftarrow$ somatorio + PESODAENTIDADE(frase, e_i)
    end for
    return $somatorio$
end function
function PESOENTREPARDEFRASES($frase_1$, $frase_2$)
    $SomatorioDePesos_1 \leftarrow$ SOMATORIODEPESOS($frase_1$)
    $SomatorioDePesos_2 \leftarrow$ SOMATORIODEPESOS($frase_2$)
    $grau_1 \leftarrow$ total de entidades na $frase_1$
    $grau_2 \leftarrow$ total de entidades na $frase_2$
    for $e_i \in frase_1$ do
        if $e_i \in frase_2$ then
            $pesoE_{i1} \leftarrow$ PESODAENTIDADE($frase_1$, e_i)
            $pesoE_{i2} \leftarrow$ PESODAENTIDADE($frase_2$, e_i)
            $Importancia_1 \leftarrow$ (pesoE$_{i1}$/SomatorioDePesos$_1$)
            $Importancia_2 \leftarrow$ (pesoE$_{i2}$/SomatorioDePesos$_2$)
            $funcaoPotuacao_1 \leftarrow$($grau_1$ / ($grau_1$ + $grau_2$)
            $funcaoPotuacao_2 \leftarrow$($grau_2$ / ($grau_1$ + $grau_2$)
            $W_{S_{ij}} \leftarrow W_{S_{i_j}}$ + (($funcaoPotuacao_1$ * $importancia_1$) + ($funcaoPotuacao_2$ * $importancia_2$))
        end if
    end for
    return $W_{s_{ij}}$
end function
$N \leftarrow$ total de frases no texto
for $frase_1 \wedge frase_2 \in frases$ do
    $SomarPesos \leftarrow$ SomarPesos + PESOENTREPARDEFRASES($frase_1$, $frase_2$)
end for
$F5 \leftarrow$SomarPesos / N
Fim

3.3.6 Transições de Entidades (**F6**, **F7**, **F8** e **F9**)

As entidades mencionadas podem ter papeis diferentes (e.g. sujeito ou objeto). Esta métrica permite quantificar cada tipo de transição de entidades entre frases adjacentes no texto (ver Algoritmo 6). Com base em [Pitler2008] e [Todirascu2013] calculou-se a probabilidade de transição sujeito-sujeito (**F6**), sujeito-objeto (**F7**), objeto-objeto (**F8**) e objeto-sujeito (**F9**) , conforme as Equações (3.12, 3.13, 3.15 e 3.14).

Algorithm 6 Pseudo código para a transição de entidades

Inicio
$listaS \leftarrow$ lista de frase no texto
for $S_i \in$ listaS do
    for $e_j \in S_i$ do
        if $(e_i \leftarrow sujeito \in S_i) \wedge (e_i \leftarrow sujeito \in S_{i+1})$ then
            $ss \leftarrow ss + 1$
            $soma \leftarrow soma + 1$
        else if $(e_i \leftarrow sujeito \in S_i) \wedge (e_i \leftarrow objeto \in S_{i+1})$ then
            $so \leftarrow so + 1$
            $soma \leftarrow soma + 1$
        else if $(e_i \leftarrow objeto \in S_i) \wedge (e_i \leftarrow objeto \in S_{i+1})$ then
            $oo \leftarrow oo + 1$
            $soma \leftarrow soma + 1$
        else if $(e_i \leftarrow objeto \in S_i) \wedge (e_i \leftarrow sujeito \in S_{i+1})$ then
            $os \leftarrow os + 1$
            $soma \leftarrow soma + 1$
        end if
    end for
end for
$F6 \leftarrow ss/soma$
$F7 \leftarrow so/soma$
$F8 \leftarrow oo/soma$
$F9 \leftarrow os/soma$
Fim

$$F6 = \frac{1}{N}\sum_{i=0} ss_i \quad (3.12)$$

$$F7 = \frac{1}{N}\sum_{i=0} so_i \quad (3.13)$$

$$F8 = \frac{1}{N}\sum_{i=0} oo_i \quad (3.14)$$

$$F9 = \frac{1}{N}\sum_{i=0} os_i \quad (3.15)$$

Onde ss_i é a transição sujeito-sujeito, so_i a transição sujeito-objeto, oo_i a transição objeto-objeto, os_i a transição objeto-sujeito e N o total de ocorrências de transições no texto.
Para esta métrica é necessário efetuar a $sentence\ segmentation$ do texto e, em cada uma das frases resultantes, fazer a etiquetagem morfo-sintática completa, a fim de se determinar as entidades presentes e o papel sintático que têm nas frases (S_i e S_{i+1}). Neste trabalho, nos textos em Português o processo foi concretizado através da anotação do texto pelo VISL (ver Secção 3.4.4) e em Inglês recorreu-se às funções da biblioteca *Stanford CoreNLP* (ver Secção 3.4.3).
Por exemplo, uma transição F6 ocorre quando uma entidade presente na frase S_i como sujeito, aparece também na frase S_{i+1} também como sujeito. O mesmo argumento é válido para outros tipos de transições.

O resultado de cada uma dessas medidas é um valor em [0, 1.00] que mede a força média de uma determinada transição.

3.3.7 Operadores Lógicos por Frase (**F10**)

Esta métrica [Graesser2004] é de simples computação e consiste na obtenção da média de operadores lógicos (e.g. ”e”, ”ou”, ”não”e ”se ... então”) nas frases. Gramaticalmente os operadores (ou conectivos) lógicos servem para ligarem partes do discurso (e.g. orações e frases). Apesar de muitas vezes serem utilizados como recurso estilísticos (polissíndeto), os níveis altos desses operadores podem tornar as frases mais complexas, exigindo a um maior esforço mental por parte do leitor e, este facto, pode atrasar o tempo de leitura de um texto; daí a necessidade de se verificar se realmente existe uma diferença entre os TExcel (ver Secção 4.1.1) e TMeds (ver Secção 4.1.2). Para quantificar esse marcador do discurso recorreu-se Equação 3.16.

$$F10 = \frac{1}{N}\sum_{i=1}^{N} X_{OL}\left(w_i\right) \quad \textbf{(3.16)}$$

Com X_{OL} a função característica para OL (conjunto de operadores lógicos), i.e.,

$$X_{OL}(w) = \begin{cases} 1 \; se \; w \; \in \; OL \\ 0 \; se \; w \; \notin \; OL \end{cases}$$

e N é o comprimento do texto.

Exemplo:

“Se o país reduzir as formalidades burocráticas e o nível de desconfiança nas instituições públicas, eliminar obstáculos de infraestrutura e as ineficiências no trânsito de mercadorias e ampliar a publicação de informações envolvendo exportação e importação, então o país reduzirá o custo do comércio exterior”. (casa das questões)[9]

Neste exemplo verifica-se que os conectivos lógicos (”se”, ”e” e ”então”) exercem um papel sintático importante nas frases. Mas, por serem demasiados, obrigam a um esforço mental adicional para a compreensão do enunciado. A aplicação de **F10** resultaria ao valor 0,133 (i.e. 6/45).

3.3.8 Densidade de Pronomes (**F11**)

Os pronomes são elementos muito importantes na coerência do texto. Muitas vezes são utilizados para substituir as entidades mencionadas para evitar a repetição, para além de servirem de recursos para se estabelecer uma cadeia de referências anafóricas. A sua distribuição nas frases podem ser quantificadas em termos de proporção ou incidência. A aplicação dessa medida foi

[9] https://acasadasquestoes.com.br/simulado/raciocinio-logico/conectivo-se-entao-condicional (pesquisada em 5/07/2018).

inspirada em [Todirascu2013] e para a sua quantificação utilizou-se a Equação 3.17.

$$F11 = \frac{1}{N}\sum_{i=1}^{N} X_{Pro}(w_i) \quad (3.17)$$

Onde X_{Pro} é a função característica para Pro (conjunto de pronomes) i.e.,

$$X_{Pro}(w) = \begin{cases} 1 \; se \; w \; \in \; Pro \\ 0 \; se \; w \; \notin \; Pro \end{cases}$$

e N é o comprimento do texto.

A interpretação do seu resultado é que quanto mais próximo de 1 estiver, maior é a densidade de pronomes no texto e mais próximo de 0 a densidade é baixa.

3.3.9 Sobreposição de Substantivos (**F12**)

Muitas vezes a distribuição de substantivos no texto contribui para os níveis de coesão entre as frases. Esta métrica ([Crossley2016], [Todirascu2013] e [Pitler2008]) visa quantificar a sobreposição dos substantivos em frases adjacentes (ver Algoritmo 7). Por ser calculado em cada par de frases, este marcador traduz a coesão local do texto. A sua aplicação baseia-se na Equação 3.18.

$$F12 = \frac{1}{N-1} \sum_{\substack{i=1 \\ w \; \in \; Sub}}^{N-1} \left(\frac{t(w, S_i \oplus S_{i+1})}{t(w, S_i) + t(w, S_{i+1})} \right) \quad (3.18)$$

Onde $p(w, S_i \oplus S_{i+1})$ é o total de substantivos w que pertencem nas frases S_i e S_{i+1}, $t(w, S_i)$ o total de substantivos w na frase S_i, N o total de frases no texto e Sub o conjunto de substantivos.

Algorithm 7 Pseudo código para o calculo da sobreposição de substantivos

```
1: Inicio
2: listaS ← lista de frases no texto
3: N ← | listaS |
4: for S_i ∈ frases do
5:     if S_i¬ (última frase) then
6:         lista1 ← lista de substantivos na frase S_i
7:         lista2 ← lista de substantivos na frase S_{i+1}
8:         total1 ← número de substantivos na frase S_i
9:         total2 ← número de substantivos na frase S_{i+1}
10:        for w_i ∈ lista1 do
11:            if w_i ∈ lista2 then
12:                contador ← contador +1
13:            end if
14:        end for
15:        soma ← soma + (contador / (total1 + total2))
16:    end if
17: end for
18: F12 ← soma / N - 1
19: Return F12
20: textbfFim
```

3.3.10 Média de Pronomes por Frase (**F13**)

Esta métrica visou quantificar a incidência de pronomes por cada frase do texto (ver Algoritmo 8). Por ser obtido ao nível da frase, é considerada como marcador de coesão local. O seu calculo baseou-se na equação 3.19.

$$F13 = \frac{1}{N} \sum_{\substack{i=1 \\ p \in Pro}}^{N} \left(\frac{t(p, S_i)}{n} \right) \quad (3.19)$$

Onde $t(p, S_i)$ é o total de pronomes p na frase S_i, n o total de palavras na frase S_i, Pro o conjunto de pronomes e N o número de frases no texto.

Algorithm 8 Pseudo código para o calculo da média de pronomes

```
1: Inicio
2: listaS ← lista de frases no texto
3: N ← | listaS |
4: for Si ∈ listaS do
5:     n ← total de palavras em Si
6:     tp ← total de pronomes na frase Si
7:     soma ← soma + (tp / n)
8: end for
9: F13 ← (1 /N) * soma
10: Return F13
11: Fim
```

3.3.11 Similaridade entre Parágrafos Iniciais e Finais (**F14**)

Muitas vezes os primeiros e últimos parágrafos de um texto refletem o tema abordado. Este pressuposto pode permitir que exista uma alta similaridade entre essas partes do texto. Para se obter a similaridade entre parágrafos iniciais e finais foi usada a função *similarity* da biblioteca *hultiglib* pertencente à UBI (ver Algoritmo 9).

Algorithm 9 Pseudo código para o calculo da similaridade entre parágrafos iniciais e finais

```
1: Inicio
2: listaS ← lista de frases no texto
3: N ← | listaS |
4: if N ≥ 3 then
5:     β ← N/3
6:     if β ≥ 3 then
7:         for Si ∈ { 0 to β } do
8:             frasesIni ← frasesIni ⊕ Si
9:         end for
10:        for i ∈ { 2β to N } do
11:            frasesFini ← frasesFini ⊕ Si
12:        end for
13:    end if
14: end if
15: F14 ← σ(frasesIni, frasesFini)
16: Return F14
17: Fim
```

Esta função é a implementação de um algoritmo baseado na equação 3.20 e calcula semelhança lexical entre dois textos, com base em evidências locais. Apoiando-se na lei de $Zipf$, considera

a frequência relativa e o comprimento da palavra e faz uma adaptação do vetor TF * IDF, sem computar o peso do IDF.

$$\sigma(t_1, t_2) = \frac{\sum\limits_{w \in V(t_1) \cap V(t_2)} \left[\left(\frac{|w|}{P(w|t_1)} \right) * \left(\frac{|w|}{P(w|t_2)} \right) \right]}{\sqrt{\sum\limits_{w \in V(t_1)} \left(\frac{|w|}{P(w|t_1)} \right)^2 * \sum\limits_{w \in V(t_2)} \left(\frac{|w|}{p(w|t_2)} \right)^2}} \quad (3.20)$$

Onde $\sigma(t_1, t_2)$ é a função similaridade entre os textos t_*, $|w|$ é o comprimento da palavra w, $P(w|t_*)$ é a probabilidade da palavra w no texto t_* e V_t o vocabulário do texto t.

3.3.12 Similaridade entre Frases (**F15**)

Esta métrica foi utilizada para computar a similaridade lexical entre as frases no texto. Tal como na métrica 3.3.11, também recorreu-se a função *similarity* da biblioteca *hultiglib* (ver Algoritmo 10). Para tal, o texto foi segmentado em frases e calculou-se os pesos da similaridade entre par de frases adjacentes. O valor final é a média aritmética desses pesos, conforme ilustrado na equação 3.21.

$$F15 = \frac{1}{N-1} \sum_{i=1}^{N-1} \sigma(S_i, S_{S+1}) \quad (3.21)$$

Onde $\sigma(S_i, S_{i+1})$ é a função similaridade entre as frases S_i e S_{i+1} e N é o número de frases no texto. Freitas [Freitas2013] considera como métrica de coerência global por ser calculada ao longo do texto de forma sequencial.

Algorithm 10 Pseudo código para o calculo da similaridade entre frases

```
Inicio
listaS ← lista de frases no texto
N ← | listaS |
if N > 1 then
    for S_i ∈ { 0 to N-1 } do
        soma ← soma + σ(S_i, S_{i+1})
    end for
end if
F15 ← (1 / (N-1) ) * soma
Return F15
Fim
```

3.3.13 Sobreposição Média de Palavras (**F16**)

Esta métrica permite detetar a incidência da sobreposição de palavras entre frases adjacentes (ver Algoritmo 11). Consiste na segmentação do texto em frases e em *tokens*. Seguidamente calcula os índices de palavras sobrepostas em cada par de frases. Este cálculo é feito com base a Equação 3.22 e, por ser feito sequencialmente em todo texto, é considerado de métrica para coesão geral.

$$F16 = \frac{1}{N} \sum_{i=1}^{N-1} \left(\frac{t(w, S_i \oplus S_{i+1})}{t(w, S_i) + t(w, S_{i+1})} \right) \quad (3.22)$$

Onde $t(w, S_i \oplus S_{i+1})$ é o total de *tokens* w que pertencem nas frases S_*, $t(w, S_*)$ é o total de *tokens* w na frase S_* e N é o numero de frases no texto.

Algorithm 11 Pseudo código para o calculo da sobreposição de palavras

```
1: Inicio
2: listaS ← lista de frases no texto
3: N ← | listaS |
4: for S_i ∈ { 0 to N-1 } do
5:     vocab ← vocabulário de tokens das frases S_i e S_{i+1}
6:     totalTokens ← | vocab |
7:     for w_i ∈ vocab do
8:         if w_i ∈ S_i ∧ w_i ∈ S_{i+1} then
9:             contador ← contador +1
10:        end if
11:    end for
12:    soma ← soma + (contador / totalTokens)
13: end for
14: F16 ← (1 / N) * soma
15: Return F16
16: Fim
```

3.3.14 Densidade de Pronomes por Substantivos (**F17**)

Esta métrica permite saber a proporção de pronomes por substantivos no texto (ver Algoritmo 12). Consiste na segmentação do texto em frases e *tokens*, seguida da análise sintática (*parser*) de cada frase. Depois do *parser* faz-se o cálculo da **F17** baseado na Equação 3.23.

$$F17 = \frac{1}{N} \sum_{i=1}^{N-1} \left(\frac{t(p, S_i)}{t(w, S_i)} \right) \quad (3.23)$$

Onde $t(p, S_i)$ é o total de pronomes p na frase S_i , $t(w, S_i)$ é o total de substantivos w na frase S_i e N é o numero de frases no texto.

Algorithm 12 Pseudo código para o calculo da densidade de pronomes por substantivos

```
1: Inicio
2: listaS ← lista de frases no texto
3: N ← | S |
4: for S_i ∈ listaS do
5:     totalDePronomes ← número de pronomes na frase S_i
6:     totalDeSubstantivos ← número de substantivos na frase S_i
7:     somatorio ← somatorio + (totalDePronomes / totalDeSubstantivos)
8: end for
9: F17 ← (1/N) * somatório
10: Return F17
11: Fim
```

3.3.15 Diversidade de Palavras de Conteúdo (**F18**)

Consiste na segmentação do texto em frases e na análise sintática de cada uma. Tem por objetivo detetar a variação de palavras de conteúdo distribuídas no texto. Por ser calculada ao nível do texto todo, é considerada como métrica de coerência global. Tal como a métrica 3.3.1, o seu

cálculo baseia-se na Equação 3.2 e para o MTLD normal e reverso (ver Equação 3.1) utilizam-se apenas as palavras de conteúdo. Para a computação dessa métrica recorreu-se no Algoritmo 1.

3.3.16 Comprimento Médio da Palavra (**F19**)

Esta métrica básica e de fácil computação (ver Algoritmo 13), permite detetar a média do comprimento das palavras. Consiste na *tokenização* do texto e na obtenção do peso (comprimento) de cada palavra. O resultado é a média aritmética de todos os pesos, conforme a Equação 3.25.

$$F19 = \frac{1}{N}\sum_{i=1}^{N-1} l(w_i) \quad \textbf{(3.24)}$$

Onde $l(w_*)$ é o comprimento do termo w_* e N é o total de palavras no texto.

Algorithm 13 Pseudo código para o calculo do comprimento médio da palavra

```
Inicio
palavras ← lista de palavras no texto
N ← | palavras |
for w_i ∈ palavras do
    peso ← | w_i |
    soma ← soma + peso
end for
F19 ← (1/N) * soma
Return F19
Fim
```

3.3.17 Comprimento Médio da Frase (**F20**)

A computação desta métrica é semelhante a 3.3.16, com a diferença de utilizar frases em vez de palavras. Consiste na segmentação do texto em frases e, em cada uma, obter o seu comprimento. O resultado é a média aritmética do somatório do comprimento de cada frase dividido pelo total de frases no texto. O seu cálculo pode ser feito com base na Equação 3.25,

$$F20 = \frac{1}{N}\sum_{i=1}^{N-1} l(S_i) \quad \textbf{(3.25)}$$

Onde, $l(S_i)$ é o comprimento da frase S_i e N é o total de frases no texto.

3.3.18 Deteção de Anáfora (**F21**)

Esta métrica permite detetar a distribuição de anáforas no texto (ver Algoritmo 14). Consiste na segmentação do texto em frases e na *tokenização* destas, seguida da análise de ocorrências de termos anafóricos em orações da mesma frase e no princípio de todas as frases. O seu calculo baseou-se na Equação 3.26.

$$F21 = \left(\sum_{i=1}^{N} \frac{l(ta, S_i)}{|S_i|} * \frac{1}{n}\right) + \frac{\sum_{i=1}^{N} X_i}{N} \quad \textbf{(3.26)}$$

Onde $l(ta, S_i)$ é o total de orações com termos anafóricos ta na frase S_i, $|S_i|$ é o comprimento de S_i, n é o número de frases com $l(ta, S_i) > 1$, X_i é cada uma das frases com termos anafóricos ta no princípio e N é o total de frases no texto.

Algorithm 14 Pseudo código para o calculo do nível de anáforas

```
1: Inicio
2: listaS ← lista de frases no texto
3: N ← | listaS |
4: function Anaforas(tipo)
5:     if tipo ← orações then
6:         for S_i ∈ listaS do
7:             comprimento ← | S_i |
8:             oracoes ← lista de orações da frase S_i
9:             for t_i ∈ orações do
10:                w_1 ← primeirapalavradet_i
11:                if freq(w_1) > 1, w_1 ∈ S_i then
12:                    total ←total + 1
13:                    for t_j ∈ orações do
14:                        if (t_i ≠ t_j) ∧ (w_1 é 1ª palavra ∈ t_j) then
15:                            contador ← contador + 1
16:                        end if
17:                    end for
18:                end if
19:            end for
20:            soma ← soma + (contador / comprimento)
21:        end for
22:        ocorrencias ← soma * (1 / total)
23:    else if tipo ← frases then
24:        for S_i ∈listaS do
25:            w_1 ← 1º termo em S_i
26:            if w_1 é 1ª palavra ∈ S_{i+1} then
27:                ocorrencias ← ocorrencias + 1
28:            end if
29:        end for
30:    end if
31:    Return ocorrencia
32: end function
33: F21 ← (Anaforas(oracoes) + Anaforas(frases)) / N
34: Fim
```

3.4 Recursos Utilizados

3.4.1 HultigLib

A *HultigLib* é uma biblioteca escrita em Java que permite executar tarefas de processamento de texto. O projeto de desenvolvimento dessa biblioteca pertence ao *Human Language Technology Information Group* (HULTIG) [10] e o seu código fonte está disponível sob os termos da Licença Pública Geral (GPL).

A *HultigLib* tem integrada a *openNLP* (Ver Secção 3.4.2) e permite, de forma eficaz, o processamento de grandes coleções de textos, para uma variedade de aplicações (por exemplo

[10]Grupo de investigação do Departamento de Informática, da Universidade da Beira Interior. Está focalizado em trabalhos relacionados com o processamento automático da linguagem humana. Para mais informações aceder ao endereço http://hultig.di.ubi.pt

a identificação de paráfrase em corpora). A *Text* (lista de elementos da *Sentence*), *Sentence* (lista de elementos da *Word*) e a *Word*, são as suas principais funções. Têm sido utilizadas para diversos trabalhos de pesquisa (por exemplo na detenção de plágio). Neste trabalho formam utilizadas maior parte das funções dessas três classes nas diversas tarefas de processamento do texto (por exemplo, a extração do texto em ficheiros, a obtenção de *tokens* e do vocabulário do texto e a similaridade lexical).

3.4.2 OpenNLP

A *Apache OpenNLP* é uma biblioteca feita em Java baseada em aprendizado de maquina usada em PLN. Permite a execução de várias tarefas de processamento do texto como a segmentação de sentença, *tokenização*, análise sintática, etiquetagem morfossintática, reconhecimento de entidades nomeadas, extração de sintagmas, resolução de correferências.
Este recurso está disponível em várias versões. Neste trabalho utilizou-se a versão 1.8 para aplicação de maior parte das métricas selecionadas.

3.4.3 Stanford CoreNLP

A *Stanford CoreNLP* [Manning2014] é uma biblioteca feita em Java, composta por um conjunto de ferramentas da Universidade de Stanford.

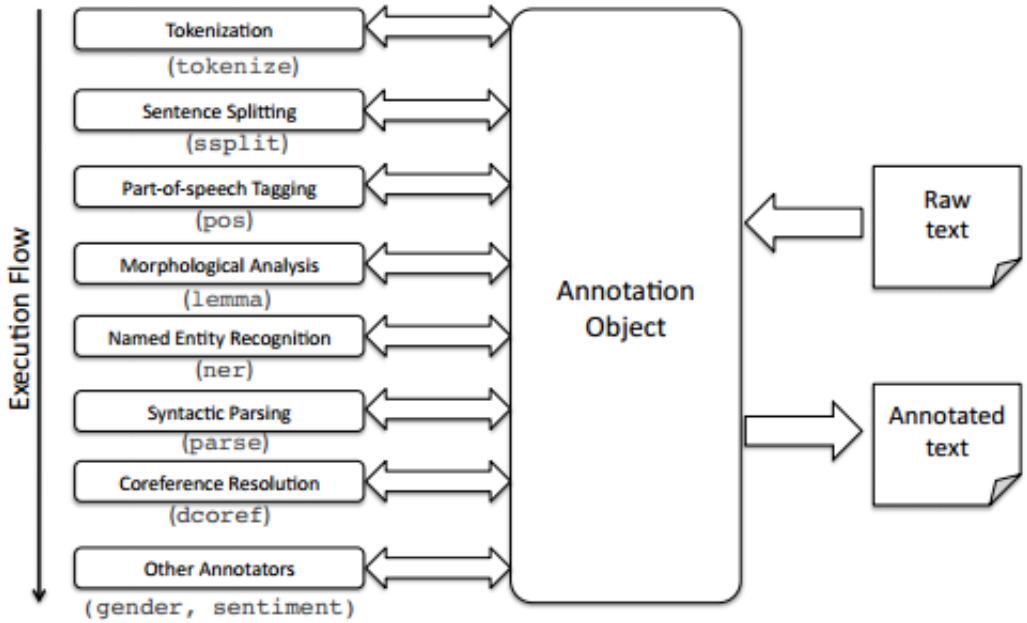

Figura 3.9: Arquitetura geral da *Stanford CoreNLP*.
Fonte: Manning et all [Manning2014].

Como ilustrado na sua arquitetura (3.9), a *Stanford CoreNLP* permite realizar várias tarefas de processamento de texto como: *tokenização*, segmentação de sentenças, marcação de partes da fala, análise morfológica, reconhecimento de entidades nomeadas, *parsing* sintático, resolução de coreferência, análise do género do texto e análise de sentimento.
Esta ferramenta está disponível sob licença GNL em muitas versões. Neste trabalho, utilizou-se a versão 3.9.1 para análise sintática do texto. O principal objetivo foi a obtenção de unidades gramaticais como sujeito e objeto nos textos em Inglês, tanto nos TMeds (Secção 4.1.2), como nos TExcel (Secção 4.1.1).

3.4.4 Visual Interactive Syntax Learning

O VISL é um sistema *online* composto por um conjunto de ferramentas e base de dados linguísticos, lançado como projeto de pesquisa e ensino na Universidade do Sul da Dinamarca em 1996. Baseou-se no sistema PALAVRAS[11] como modelo para outras línguas e, até ao momento de escrita deste trabalho, suporta a análise sintática da frase em 13 (para alguns 14, através do Bokmål norueguês) línguas diferentes (Dinamarquês, Holandês, Inglês, Português, Esperanto,Gronelandês, Francês, Alemão, Espanhol, Japonês, Italiano, Norueguês e Sueco). Para além da análise sintática, possui módulos para análise semântica, tradução automática, bem como a coleção e etiquetagem de corpora.
Neste trabalho, foi utilizado o módulo *full morphosyntactic parse*, para anotação dos textos em Português. Do texto anotado, foi possível obter elementos da frase como sujeito e objeto, para aplicação da Métrica F5 (Ver Secção 3.3.5).
Por exemplo, ao anotar a frase "O João foi buscar o seu carro ontem", tem-se o seguinte resultado:

```
o [o] <artd> DET M S @>N
João [João] <hum> PROP M S @SUBJ>
foi [ir] <fmc> V PS 3S IND VFIN @FAUX
buscar [buscar] <vt> V INF @IMV @ICL-AUX<
o [o] <artd> DET M S @>N
seu [seu] <poss 3S> <si> DET M S @>N
carro [carro] <V> N M S @<ACC
ontem [ontem] <atemp> ADV @<ADVL
```

Figura 3.10: Texto resultado da anotação da frase "O João foi buscar o seu carro ontem"pelo VISL.

Pode-se notar que "João" é o sujeito e "carro" o objeto. A identificação dessas unidades lexicais no VISL é feita através das *tags* **@SUBJ** — que indica o sujeito da frase — e **@ACC**, **@DAT** e **@PIV** — que indicam o objeto da frase. A diferença entre essa três últimas *tags* reside em **@ACC** ser usados para detetar o objeto direto, **@DAT** para o objeto direto (pronominal) e **@PIV** para detetar o objeto indireto.

3.4.5 New General Service List (NGSL)

A Nova Lista Geral de Serviços é uma lista do vocabulário de alta frequência do Inglês, derivada de 273 milhões de palavras selecionadas, dos 2 biliões de palavras da *Cambridge English Corpus* (CEC)[12], publicada em março de 2013 [Browne2014]. Contém aproximadamente 2.800 palavras e é uma atualização da anterior lista *General Service List* (GSL)[13] publicada por de Michael West em 1953. A NGSL representa aproximadamente 92% de cobertura para a maioria dos textos em Inglês geral (a mais alta de todas as listas de palavras inglesas gerais derivadas de corpus até hoje).

[11] *Parser* para a língua portuguesa que possui uma série de recursos, tais como *POS-tagging*, anotação semântica, sintática, entre outras. Encontra-se integrado no projeto VISL.

[12] Corpus de biliões de palavras do Ingles escrito e falado (Britânico e Americano), de várias fontes diferent. Para mais informações consultar o endereço: http://www.cambridge.org/about-us/what-we-do/cambridge-english-corpus

[13] Lista de aproximadamente 2.000 palavras, retiradas de um corpus de inglês escrito, para representar as palavras mais frequentes. Foi publicadas por Michael West em 1953. Para mais informações consultar o endereço: https://en.wikipedia.org/wiki/General_Service_List.

Nesse trabalho foi utilizada para a obtenção do nível de sofisticação do vocabulário do texto (ver Secção 3.3.3). Por conter maior parte das palavras de alta frequência no Inglês, constitui recurso importante a ser aplicado em métricas para avaliação de escrita. Por exemplo, uma palavra é considerada sofisticada quando não integra essa lista e não é uma gíria ou calão.

3.4.6 TestH

A *TestH* é uma biblioteca ANSI C desenvolvida para análise de fractais. É resultado de um projeto (ainda em andamento) levado a cabo por investigadores da Universidade da Beira Interior, para atender a trabalhos de investigação científica relacionadas com estruturas auto-semelhantes. Segundo os seus autores, o objetivo principal é fornecer aos investigadores os meios para estimar o parâmetro de Hurst e gerar sequências com a propriedade de auto-semelhança incorporada nos dados, podendo ser usada em uma ampla variedade de campos de pesquisa(Fernandes et all [Fernandes2014]).
No nosso trabalho utilizamos a versão mais recente dessa biblioteca, cujo código fonte pode ser encontrado no endereço https://github.com/diogo-fernan/testh.
A sua API bastante amigável permite a leitura de arquivos com dados numéricos, computar a sua informação estatística e, através dos estimadores do parâmetro (ou expoente) de Hurst, imprimir o resultado do processamento (ver Secção 4.2.1).

3.4.7 Outros Recursos

Para além dos já mencionados, foram ainda utilizados dicionários disponibilizados pelo projeto *OpenOffice* (em Português e em Inglês). A escolha desses dicionários deveu-se ao facto de serem os mais usados em *softwares* de aplicação conhecidos (por exemplo o *Office*, *LibreOffice* e o *TextStudio*).
Por não existir um projeto semelhante ao NGSL em Português, recorreu-se a lista de palavras comuns da língua portuguesa disponibilizada pelo site *quizlet.com*. Essa lista foi utilizada para os mesmos fins da NGSL para textos em Português.

3.5 Sumário

Neste capítulo abordou-se o método de avaliação da qualidade e estética do texto proposto. Inicialmente foi apresentada a metodologia adotada para a pesquisa, em torno da qual baseou-se todo trabalho realizado.
Seguidamente descreveram-se os marcadores do texto selecionados e as respetivas métricas envolvidas, assim como os algoritmos utilizados nas experiências e, finalmente, os principais recursos utilizados para aplicação das métricas selecionadas.
No capítulo 4 serão descritas as experiências realizadas, bem como os resultados obtidos e a sua respetiva discussão.

Capítulo 4

Experiências e Resultados

Este capítulo aborda as experiências realizadas por meio da metodologia descrita anteriormente e os resultados obtidos na avaliação da qualidade e estética do texto. Será apresentada a descrição pormenorizada dos corpora utilizado e o ambiente de teste, assim como dos resultados obtidos. A sua descrição tem a seguinte estrutura: **Secção 4.1** — Conjunto de Dados; **Secção 4.2** — Testes e Resultados; **Secção 4.3** — Discussão dos Resultados e **Secção 4.4** — Sumário.

4.1 Conjunto de Dados

Como já salientado na Secção 3.1, para as experiências, foi constituído um corpora composto por TExcel e outro de TMeds, alguns escritos em Português e outros em Inglês. No total foram selecionados 110 ficheiros (entre livros, redações e textos sobre temas variados). A Tabela 4.1 mostra a distribuição dos dois conjuntos.

	Corpora					
	TExcel			**TMeds**		
	PT	**EN**	**Total**	**PT**	**EN**	**Total**
Ficheiros	20	30	50	20	40	60
Frases	-	-	497.360	-	-	532.680
Tokens	-	-	5.795.984	-	-	4.914.775
Palavras	-	-	4.751.472	-	-	3.881.223

Tabela 4.1: Distribuição do Corpora Utilizado.

4.1.1 Corpus de TExcel

O corpus aqui designado por TExcel é um conjunto de textos considerados de alta qualidade. Foram selecionadas 50 obras de diversos autores e géneros, o que totalizou 497.360 frases e 4.751.472 palavras (ver Tabela 4.1). A descrição das obras está detalhada na Tabela 4.2. Algumas delas foram escritos por autores congratulados com o *Prémio Nobel da literatura*[1] (principalmente para textos em Inglês) e outros com o *Prémio Oceano de Literatura*[2] (principalmente para textos em Português). Consideramos textos de excelente qualidade por pertencerem a autores consagrados e com experiência acumulada há anos na escrita. os textos em Português foram anotados no VISL (Ver Secção 3.4.4) para o devido processamento.

[1] Atribuído anualmente a escritores com pesquisas de grande relevância na área de literatura. Para mais informações consultar o endereço https://www.nobelprize.org/

[2] É um dos prémios literários mais importantes entre os países de língua portuguesa. Até 2014 foi conhecido como Prémio Portugal Telecom de Literatura. Para mais informações consultar o endereço https://pt.wikipedia.org/wiki/Oceanos-Prêmio_de_Literatura_em_Língua_Portuguesa

Nº	Titulo	Autor	Género
1	A Máquina de Fazer Espanhóis	Valter Hugo Mãe	Ficção
2	Comissão das Lágrimas	António Lobo Antunes	Conto
3	Ensaio sobre a Cegueira	José Saramago	Ficção (Romance)
4	Os céus de Judas	António Lobo Antunes	Ficção
5	Explicações dos Pássaros	António Lobo Antunes	Ficção (Romance)
6	Never Let Me Go	Kazuo Ishiguro	Ficção científica
7	Seeing	José Saramago	Ficção (romance)
8	As naus	António Lobo Antunes	Ficção
9	The Buried Giant	Kazuo Ishiguro	Fantasia (Romance)
10	The Remains of the Day	Kazuo Ishiguro	Ficção (História)
11	A guerra não tem rosto de mulher	Svetlana Alexijevich	Narrativa
12	Vida Querida	Alice Munro	Romance
13	When We Were Orphans	Kazuo Ishiguro	Ficção policial (Romance)
14	All the Names	José Saramago	Ficção (Romance)
15	Away from Her	Alice Munro	Ficção
16	Baltasar and Blimunda	José Saramago	Romance
17	Dance of the Happy Shades	Alice Munro	Ficção
18	Matteo perdeu o emprego	Gonçalo Manuel de Albuquerque Tavares	Romance
19	Death with Interruptions	José Saramago	Ficção
20	Death at Intervals	José Saramago	Ficção
21	Family Furnishings	Alice Munro	Ficção literária
22	Flores da Ruína	Patrick Modiano	Ficção
23	Friend of My Youtho	Alice Munro	Ficção
24	Jerusalém	Gonçalo Manuel de Albuquerque Tavares	Romance
25	Journey to Portugal	José Saramago	Literatura de viagem
26	Julieta	Alice Munro	História
27	Lives of Girls and Women	Alice Munro	História
28	The Love of a Good Woman	Alice Munro	Ficção (Antologia)
29	O beijo na nuca	Dalton Trevisan	Contos
30	Manual of Painting and Calligraphy	José Saramago	Ficção (Romance)
31	The Moons of Jupiter	Alice Munro	Conto
32	O fim do homem soviético	Svetlana Alexijevich	Biografia
33	Open Secrets	Alice Munro	Contos
34	Raised from the Ground	José Saramago	Romance
35	Runaway	Alice Munro	Contos
36	Contemporary Critical Perspectives	Kazuo Ishiguro	-
37	Selected Stories	Alice Munro	Biografia
38	Skylight	José Saramago	Romance
39	Small Memories	José Saramago	Biografia
40	Snow	Orhan Pamuk	Ficção
41	Something I've Been Meaning to Tell You	Alice Munro	Contos
42	O Conto da Ilha Desconhecida	José Saramago	Conto
43	Eu Hei-de Amar Uma Pedra	António Lobo Antunes	Romance
44	The Cave	José Saramago	Romance
45	A Viagem do Elefante	José Saramago	História-Romance
46	O Evangelho segundo Jesus Cristo	José Saramago	História-Romance
47	História do Cerco de Lisboa	José Saramago	História-Romance
48	O Sonâmbulo Amador	José Luiz Passos	Romance
49	The Progress of Love	Alice Munro	Ficção
50	A Jangada de Pedra	José Saramago	Romance

Tabela 4.2: Descrição detalhada das obras utilizadas na constituição do corpus de TExcel.

4.1.2 Corpus de TMeds

Para a constituição desse corpus, composto por textos considerado de baixa qualidade, foi coletado um conjunto de redações da *VOL*[3] e de *Blogs*[4] [Koppel2006]. Os textos recolhidos da *VOL* são todos em Português e foram anotados no VISL (ver Secção 3.4.4) para permitirem a análise morfo-sintática, como apresentado na Secção 3.1. Esses textos são considerados de qualidade reduzida por dois motivos principais:

a) Por um lado lado, os textos em Inglês foram escritos por *internautas* de várias idades, sem a mínima preocupação dos aspetos linguísticos necessários a uma boa escrita. Não houve a intervenção humana para avaliar a sua qualidade e filtrar os melhores.

b) Por outro, os textos em Português foram escritos por *internautas* e, posteriormente, submetidos a base de dados de redações da *VOL* para serem avaliados por peritos humanos (e.g. linguistas e professores), o que é bastante valioso para as configurações adotadas no presente trabalho. Deste último grupo, selecionaram-se aqueles que tiveram uma baixa classificação. A única correção feita aos originais foi a dos erros ortográficos e adequação ao novo acordo ortográfico.

No total foram selecionados 60 ficheiros de tamanho variado, onde, 33,3% foram escritos em Português e 66,7% em Inglês. Este corpus forneceu 532.680 frases e 3.881.223 palavras utilizadas para os testes (ver Tabela 4.1).
Aqui não foi apresentada uma tabela detalhada deste corpus por não se tratar de obras literárias divulgadas e não haver necessidade para tal. Por outro lado, alguns autores (e.g. da base de redação da *VOL*) preferem manter o anonimato quando submetem as suas redações para efeitos de avaliação.

4.2 Testes e Resultados

Todas as experiências propostas foram realizadas por meio de uma aplicação em Java, construida para esse propósito. O código fonte está disponível[5] no *GitHub* para consultas e eventuais atualizações futuras.

Computador utilizado	
Marca	HP
Modelo	ProBook 6570b
Processador	Intel Core i3
RAM	8 GB
HD	Hd 500gb 5000rpm Bufer16mb Sata3

Tabela 4.3: Características da máquina utilizada para as experiências.

[3] Repositório de redações sobre várias temáticas, submetida a um corpo de júri para serem avaliada qualitativamente. Para mais informações consultar o endereço https://educacao.uol.com.br/bancoderedacoes/

[4] É um corpus de dezenas de milhares de *Blogs*. Contém cerca de 300 milhões de palavras e foi estudado em [Koppel2006] para descobrir diferenças significativas no estilo de escrita e conteúdo entre *internautas* do sexo masculino e feminino, bem como entre autores de diferentes idades.

[5] https://github.com/Domicardio/Qualidade-Do-Texto

Foram realizados dois tipos de experiências para avaliar a qualidade e estética do texto:

a) Análise de Auto-Semelhança;
b) Análise Geral do Texto.

As experiências feitas na primeira etapa decorreram num tempo de cinco horas, vinte e três minutos, doze segundos e cento e vinte e três décimos (5:23:12:123). Na segunda fase o tempo foi relativamente maior, isto é, dezoito horas, 47 minutos, dezanove segundos e 136 décimos (18:47:19:136).
As características da máquina utilizada (ver Tabela 4.3) foi um dos fatores que contribuiu para a morosidade no processamento. Este fator, aliado a análise da eficiência dos algoritmos utilizados constituem alguns dos requisitos a serem melhorados em trabalho futuro.

4.2.1 Análise da Auto-Semelhança

A primeira fase, visou detetar a existência de estruturas Auto-Semelhantes no texto. Para a referida análise subdividimos cada texto em blocos com tamanhos iguais. Optamos por trabalhar com blocos de 20 frases em vez de blocos de palavras, pelo facto da frase representar uma unidade lexical com sentido completo. Outra razão dessa opção é que a maior parte das métricas utilizadas no nosso trabalho, segmentam o texto em frase e só depois é que capturam os elementos específicos (e.g. pronomes, *tokens*, *types*, operadores lógicos e substantivos) para o devido processamento. A constituição dos blocos para os TMeds (Secção 4.1.2) e TExcel (Secção 4.1.1) pode ser vista na Tabela (4.4).

	Blocos de texto					
	TExcel			**TMeds**		
	PT	EN	Total	PT	EN	Total
Blocos	9.872	14.996	24.868	10.953	15.681	26.634

Tabela 4.4: Distribuição dos textos por blocos de 20 frases utilizado nas medições de Auto-Semelhança.

Para a experimentação das medidas e obtenção dos primeiros resultados, cada bloco foi considerado um texto independente. Todo processamento foi feito pela aplicação construida para o efeito e consistiu em avaliar os blocos individualmente através dos teste com as métricas selecionadas. Nessa experiência, constitui-se um conjunto de 42 ficheiros, contendo os dados numéricos resultantes das medições executadas. Desses ficheiros, 50% corresponde aos blocos de TExcel e outros 50% para os de TMeds.
Por exemplo, a medida **F1** (Secção 3.3.1) originou um ficheiro (**F1.txt**) contendo 24.868 números decimais para os blocos dos TExcel. Esses números traduzem os resultados obtidos pela aplicação da métrica **F1** nos blocos correspondentes. O mesmo procedimento foi aplicado em todos os blocos, tanto dos TExcel como dos TMeds.
Posteriormente, os 42 ficheiros resultantes dessa experiência, foram submetidos ao *TestH* (ver Secção 3.4.6) para a análise de existência de estruturas Auto-Semelhantes.
Através da API do *TestH*, é feita leitura de cada um dos arquivos de dados, processados os valores e realizadas as informações estatísticas para se estimar o parâmetro Hurst. O nível de Auto-Semelhança dos dados é obtido pelo valor do parâmetro de Hurst encontrado. O resultado detalhado dessa experiência pode ser visto na Tabela 4.5, ou graficamente na Figura 4.1.

	Parâmetro de Hurst	
	TExcel	TMeds
F1	0.4571	0.4766
F2	0.4591	0.4590
F3	0.4743	0.4657
F4	0.4682	0.4762
F5	0.4686	0.4922
F6	0.4653	0.4610
F7	0.4701	0.4633
F8	0.4861	0.4564
F9	0.4881	0.4639
F10	0.3127	0.3325
F11	0.4684	0.4825
F12	0.4648	0.4213
F13	0.4801	0.4736
F14	0.3332	0.3476
F15	0.4610	0.3548
F16	0.4661	0.4667
F17	**0.5045**	0.4749
F18	0.4737	0.4875
F19	0.4742	0.4335
F20	0.4814	0.4594
F21	0.4681	0.4891

Tabela 4.5: Descrição dos valores do parâmetro de Hurst, por características do texto, obtidos pelas experiências realizadas na análise de fractalidade.

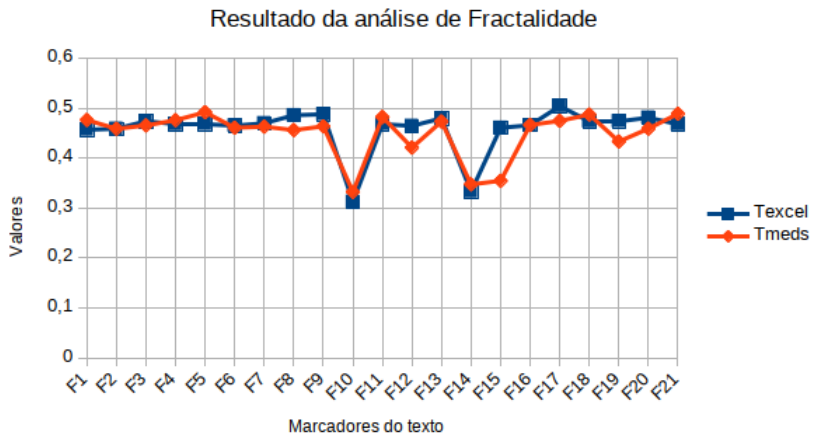

Figura 4.1: Gráfico comparativo dos resultados obtidos na análise de fractalidade nos TExel e no TMeds.

4.2.2 Análise Geral dos Textos

A segunda fase, consistiu na análise completa dos textos, i.e., na avaliação dos textos originais através da aplicação das métricas selecionadas. Cada texto forneceu um conjunto de 21 dados numéricos como resultado da aplicação das métricas propostas no presente trabalho. Esses dados serviram para se descobrir quais os atributos melhor carateriza os TExcel e se distinguem dos TMeds. O resultado das experiências realizadas nesta fase podem ser vistos com mais detalhe na Tabela 4.6 ou pelo gráfico das médias na Figura 4.2.

	Análise geral do corpora							
	TExcel				TMeds			
	min	max	$\bar{x}$	dp	min	max	$\bar{x}$	dp
F1	0,5513	0,8733	0,6756	0,0874	0,3901	1,0000	0,6638	0,1475
F2	0,7961	0,9215	0,8615	0,0311	0,0432	0,9246	0,5990	0,3241
F3	0,3241	0,8867	0,6186	0,1252	0,1399	1,0000	0,4577	0,2671
F4	0,4813	0,7136	0,5923	0,0577	0,6008	0,7326	0,6661	0,0331
F5	0,029	0,112	0,0611	0,0197	0,014	0,1218	0,0549	0,0223
F6	0,4322	0,6437	0,5280	0,0513	0,0000	1,0000	0,3939	0,2710
F7	0,0513	0,1755	0,1178	0,0285	0,0000	0,1717	0,0659	0,0560
F8	0,0591	0,2692	0,1885	0,0549	0,0000	0,4482	0,1778	0,1346
F9	0,1076	0,2207	0,1589	0,0269	0,0000	0,2727	0,0976	0,0795
F10	0,0357	0,0500	0,0428	0,0035	0,0163	0,0653	0,0407	0,0105
F11	0,0649	0,1530	0,1110	0,0216	0,0495	0,1465	0,1034	0,0208
F12	0,0127	0,0372	0,0238	0,0054	0,0045	0,0434	0,0236	0,0091
F13	0,0559	0,1280	0,0915	0,0167	0,0403	0,1101	0,0823	0,0149
F14	0,1007	0,2697	0,1769	0,0461	0,0200	0,2132	0,1157	0,0519
F15	0,0397	0,0700	0,0555	0,0065	0,0164	0,0955	0,0556	0,0188
F16	0,0406	0,5232	0,1684	0,1198	0,0000	0,3279	0,1150	0,0783
F17	0,3056	1,0000	0,6240	0,1602	0,2280	0,8965	0,5519	0,1656
F18	0,3553	1,0000	0,5696	0,1755	0,1505	0,7665	0,4140	0,1473
F19	0,6854	0,8015	0,7350	0,0300	0,6385	1,0000	0,7666	0,0932
F20	0,0001	0,0089	0,0024	0,0021	0,0000	1,0000	0,1688	0,2839
F21	0,0279	0,1018	0,0669	0,0192	0,0000	0,1667	0,0684	0,0430

Tabela 4.6: Descrição dos valores dos marcadores do texto selecionados, obtidos pelas experiências realizadas na análise geral dos corpora.

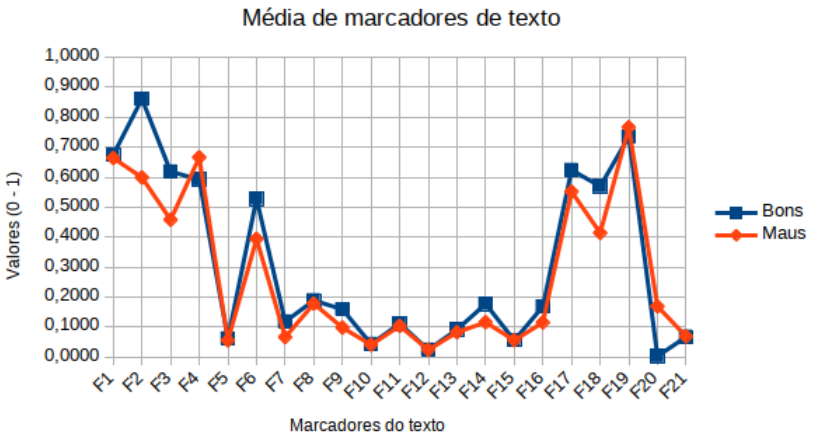

Figura 4.2: Gráfico das médias obtidas pela aplicação das métricas dos marcadores do texto selecionadas.

4.3 Discussão dos Resultados

Para aumentar a perceção dos resultados obtidos, testamos o grau de significância que as diferenças apresentam, através do teste estatístico paramétrico *t-Student* (ou teste t). Como as amostras são independentes, esse teste mostrou-se adequado para detetar diferenças significativas entre amostras dos dois grupos. Os resultados foram agrupados nas respetivas categorias como:

a) Vocabulário do texto (Tabela 4.7);
b) Coesão e coerência (Tabela 4.8);
c) Complexidade Sintática (4.9);
d) Figuras de estilo (Tabela 4.10).

4.3.1 Resultados da Primeira Experiência

A nossa primeira experiência consistiu na deteção de partículas auto-semelhantes no texto (ver Secção 4.2.1). Como já mencionado na Secção 2.5, este modelo foi sugerido em [Cordeiro2015] e, a sua utilização nesse trabalho, teve o objetivo de descobrir se pode ser usado para detetar atributos capazes de discriminarem textos da categoria TExcel (Ver secção 4.1.1) dos de TMeds (Ver secção 4.1.2).
Como podemos constatar na Tabela 4.5, dos 21 marcadores de texto selecionados, obteve-se um valor acima de 0,5 (precisamente 0,5045) na Métrica **F17** para os TExcel. Nessa experiência não foi feito o teste *t-student* devido a natureza dos valores do parâmetro de Hurst estimados. Apesar do valor de **F17** ser mais elevado em relação aos demais, esse dado não é suficientemente bom para afirmar que o marcador **F17** é característico de TExcel. O valor ideal para se afirmar com segurança a existência de auto-semelhança deve ser acima de 0.8.
Apesar do modelo não detetar nenhum marcador discriminador de qualidade e estética do texto, fica a evidência de que, com uma investigação mais profunda, o mesmo poderá fornecer dados mais confiáveis para este tipo de tarefa. No nosso trabalho, utilizamos blocos de 20 frases, em vez de blocos de palavras (como em [Cordeiro2015]). O facto do resultado ser inferior ao esperado, pode estar condicionado pelas configurações implementadas e, num trabalho futuro, deve ser explorado melhor o modelo para verificar se é possível obter melhores resultados na avaliação da qualidade e estética do texto, usando análise de auto-semelhança.

4.3.2 Resultados da Segunda Experiência

A nossa segunda experiência consistiu na aplicação das métricas selecionadas aos dois grupos de textos. Os resultados das Tabelas 4.7, 4.8, 4.9 e 4.10 refletem os valores obtidos nos testes paramétrico *t-Student* para variáveis independentes.

Vocabulário do Texto

O primeiro grupo (Tabelas 4.7) é de marcadores do vocabulário do texto (**F1**, **F2**, **F3** e **F4**). Nele podemos constatar que três desses marcadores (**F2**, **F3** e **F4**) foram bastante eficazes na distinção de TExcel e TMeds.

	Vocabulário			
	t	**p-value**	**graus de liberdade**	**Dif. média**
F1	0,507	0,613	91,155	0,0118
F2	6,237	**0,000**	60,383	**0,2625**
F3	4,082	**0.000**	88,980	**0,1609**
F4	-7,888	**0,000**	74,301	**0,0738**

Tabela 4.7: Descrição dos resultados do teste *t-student* para marcadores do vocabulário do texto.

Para a medida **F1**, no intervalo de confiança de 95% com 91,155 graus de liberdade, não há

evidências de que as médias da diversidade lexical medidas pela MTLD (Ver Secção 2.2.1) sejam diferentes entre os TExcel e TMeds **(t(91,155)=0,507; p>0,05)**. Apesar de Victoria Johansson [Johansson2009] afirmar que a LD pode ser utilizada na avaliação da qualidade do texto, este dado mostra que a **F1** não é um atributo que ajuda a distinguir os dois grupos de textos.

Para a **F2**, segundo o teste *t-student*, a média deste marcador é relativamente maior nos TExcel que nos TMeds **(t(60,383)=6,237; p<0,05)**. A nossa hipótese era de que a taxa de repetição fosse maior nos TMeds. Surpreendentemente este dado contrariou as nossas expetativas. Apesar disso, este marcador mostrou evidencias suficientes para distinção dos dois grupos de textos.

Relativamente ao marcador **F3**, o *t-student* independente mostrou que, em média, os TExcel apresentam maior número de palavras elaboradas em relação aos TMeds **(t(88,980)=4,082; p<0,05)**. Embrora Natalie e Jacqueline [Olinghouse2009] tenham encontrado resultados não muito satisfatórios com esta medida, este resultado valida a nossa desconfiança inicial de que escritores experientes produzem textos com maior proficiência do seu vocabulário em relação aos menos experientes.

Para a **F4**, o teste estatístico *t-Student* mostrou evidencias de que, em média, os TMeds apresentam mais palavras de conteúdo que os TExcel **(t(74,301)=-7,888; p<0,05)**. Em consonância com as descobertas de Victoria Johansson [Johansson2009], este marcador pode assim servir de discriminador de TExcel e TMeds.

Coesão e Coerência

O segundo grupo nesta experiência é formado por marcadores ligados à coesão e à coerência do texto (Ver Secção 3.2.2). Como podemos notar na Tabela 4.8 (valores em negrito), oito desses marcadores apresentaram diferenças significativas entre TExcel e TMeds. O teste *t-Student* mostrou que não existe diferenças entre as médias das métricas **F5**, **F8**, **F11**, **F12** e **F15** nos dois grupos de textos (com p>0,05).

	Coesão e coerência			
	t	**p-value**	**graus de liberdade**	**Dif. média**
F5	1,455	0,149	95	0,0062
F6	3,773	**0,000**	66,007	**0,1341**
F7	6,179	**0.000**	92,060	**0,0518**
F8	0,555	0,580	80,994	0,0107
F9	5,542	**0.000**	76,939	**0,0613**
F11	1,845	0,068	104	0,0076
F12	0,145	0,885	80,330	0,0002
F13	2,955	**0,004**	101	**0,0092**
F14	6,492	**0,000**	109	**0,0612**
F15	-0,037	0,970	66,919	0,0001
F16	2,619	**0,011**	76,695	**0,0534**
F17	2,254	**0,026**	102	**0,0721**
F18	5,036	**0,000**	107	**0,1556**

Tabela 4.8: Descrição dos resultados do teste *t-Student* para marcadores da coesão e coerência do texto.

Neste grupo, três dos cinco marcadores de coerência baseada em entidades selecionados, serviram para destingir TExcel e TMeds. O *t-Student* mostrou que, em média, as transições **S** $\rightarrow$ **S**, **S** $\rightarrow$ **O** e **O** $\rightarrow$ **O** são maiores nos TExcel que nos TMeds, com **(t(66,007)=3,773; p<0,05)**, **(t(92,060)=6,179; p<0,05)**, e **(t(76,939)=5,542; p<0,05)** respetivamente. Estes resultados também foram encontrados por Todirascu ent al [Todirascu2013] e Pitler and Nenkova [Pitler2008].

Infelizmente, as transições **O** $\rightarrow$ **S** não serviram para discriminar o dois grupos de texto. Por outro lado, ao contrário dos resultados encontrados por Mesgar and Strube [Mesgar2014], o teste *t-Student* para o valor da medida **F5**, não forneceu evidência significativa para distinguir TExcel e TMeds.

Dois dos quatro marcadores de coesão lexical permitem discriminar os dois grupos de textos. O *t-Student* mostrou que a média da similaridade lexical entre parágrafos iniciais e finais é maior nos TExcel em relação aos TMeds **($t(109)=6,492$; $p<0,05$)**. Este resultado, validou a nossa hipótese inicial de que normalmente, os escritores incorporam nos primeiros e últimos parágrafos dos seus textos, palavras relacionadas ao tema abordado. Este facto, permite uma maior semelhança lexical nessas duas secções do texto. É de salientar que a Métrica usada no cálculo do **F14** baseou-se na função *similarity* da biblioteca *HultgLib* (ver Secção 3.4.1) da UBI. O resultado obtido, permite confirmar a eficiência dessa função neste este tipo de tarefa. A **F16** também mostrou ser um bom discriminador de TExcel e TMeds. Como podemos observar na Tabela 4.8, em média os TExcel têm mais palavras sobrepostas em frases adjacentes em relação aos TMeds **($t(76,695)=2,619$; $p<0,05$)**. Este resultado está em conformidade aos dados fornecidos pela ferramenta TAACO (ver Secção 2.3) de Crosley et all [Crossley2016]. Sendo um atributo discriminador de coerência do texto, também pode ser utilizado na distinção de TExcel e TMeds.

Dois dos quatro marcadores baseados em part of speech selecionados, mostraram resultados positivos na discriminação dos dois grupos de textos. Segundo os dados obtidos pelo teste *t-Student*, em média, os TExcel apresentam uma maior proporção de pronomes por substantivos em relação aos TMeds **($t(102)=2,254$; $p<0,05$)**. Embora esse marcador não apresentou correlação significativa com a coerência do texto nos trabalho de Todirascu el all [Todirascu2013], o nosso resultado mostra que pode ser usado para diferenciar a coerência de TExcel e TMeds. Outro marcador significativo na distinção dos dos tipos de textos é o **F18**. Segundo o resultado obtido no teste *t-Student*, em média, os TExcel apresentam mais diversidade de palavras de conteúdo em relação aos TMeds **($t(107)=5,036$; $p<0,05$)**. Como já mencionado na Secção 3.3.1 a Métrica usada neste cálculo baseou-se na MTLD de McCarthy e Jarvis [McCarthy2010], com a ligeira diferença de usar apenas palavras de conteúdo, em vez de todo o texto. Embora a **F1** não tenha dado resultado positivo, a aplicação da mesma técnica com F18, resultou em valores significativos na diferenciação de TExcel e TMeds.

Complexidade Sintática

	Complexidade sintática			
	t	**p-value**	**graus de liberdade**	**Dif. média**
F10	1,462	0,148	71,375	0,0021
F19	-2,491	**0,015**	74,914	**0,0316**
F20	-4,307	**0,000**	53,007	**0,1664**

Tabela 4.9: Descrição dos resultados do teste *t-student* para marcadores da complexidade sintática.

Dos marcadores de complexidade sintática selecionados, apenas dois resultaram em valores significativos para a tarefa proposta. O resultado do *t-student* mostrou que os TMeds normalmente têm frases mais longas em comparação com os TExcel **($t(74,914)=-2,491$; $p<0,05$)** e, em média, os TMeds contêm palavras mais longas em relação aos TExcel **($t(53,007)=-4,307$; $p<0,05$)**. Segundo Cordeiro et all [Cordeiro2015], palavras mais longas transmitem mais informação que as mais curtas. Seguindo essa ideia, esperávamos que o marcador **F20** fosse característico nos TExcel. Surpreendentemente, o resultado foi totalmente diferente. Porém, esse marcador mos-

trou evidências de ser discriminatório dos dois tipos de textos e, pelas configurações das nossas experiências, demonstrou ser característico de TMeds. Por outro lado, a nossa ideia inicial era de que o marcador **F10** pudesse ajudar a discriminar TExcel e TMeds. Contra todas as nossas expetativas, os resultados do *t-Student* não mostraram evidências suficientes para essa distinção.

Figuras de Estilo

	Anáfora			
	t	**p-value**	**graus de liberdade**	**Dif. média**
F21	-0,225	0,822	84,408	0,0015

Tabela 4.10: Descrição dos resultados do teste *t-student* para marcadores de figuras de estilos.

As figuras de estilo foram pouco exploradas nas nossas experiências. Sabendo da importância que as mesmas têm nos textos e das possibilidades de serem bons discriminadores de TExcel e TMeds, deverão ser mais exploradas em trabalho futuro. No nosso trabalho apenas selecionamos a anáfora por ser de fácil comutação e por ser incorporada na métrica **F2**. Como se pode observar na Tabela 4.10, o resultado do *t-Student* para o **F21**, não foi significativo para distinguir os dois tipos de texto $\mathbf{(t(84{,}408)=-0{,}225;\ p>0{,}05)}$. Isso pressupõe dizer que os escritores menos e mais experientes incorporam (inconscientemente) os mesmos níveis de termos anafóricos nos seus textos.

4.4 Sumário

Neste capítulo abordamos todas as configurações feitas, para as experiências propostas no nosso trabalho. Foi feita a descrição pormenorizada do corpora utilizado, desde a coleta, anotação e distribuição nos dois grupos te textos. Posteriormente descrevemos o ambiente de teste e os resultados obtidos nas experiências realizadas. Finalmente foi feita a discussão dos resultados pela análise estatística paramétrica *t-Student* de variáveis independentes.
No capítulo 5 serão apresentadas as principais conclusões obtidas no nosso estudo bem como a descrição do trabalho a ser feito futuramente.

Capítulo 5

Conclusão e Trabalho Futuro

5.1 Conclusão

Nesta dissertação apresentamos o nosso método para avaliação da qualidade e estética do texto (Capítulos 3 e 4). O estudo aqui reportado teve como objetivo principal experimentar modelos matemáticos que permitam avaliar qualitativamente um texto e filtrar os marcadores mais significativos que estão na base dessa distinção. É um trabalho de natureza experimental, que cumpre o seu objetivo, ao detetar valores de marcadores com significância estatística, na distinção de textos bons e maus. Para tal, construiram-se dois corpora e aplicou-se um conjunto de 21 marcadores linguísticos para discriminar os dois tipos de textos. As experiências realizadas decorreram em duas etapas diferentes: Análise de Auto-Semelhança e Análise Geral do Texto.

Como se constatou, a primeira fase consistiu em analisar a existência de auto-semelhança através do estimador do parâmetro de Hurst e descobrir se as mesmas podem servir para distinguir a qualidade e estética do texto. Para essa experiência dividiu-se cada elemento dos corpora em blocos de 20 frases e, testou-se os 21 marcadores de texto selecionados, em cada bloco. Os dados obtidos apontam para valores inesperados para a tarefa proposta. A proporção de pronomes por substantivos foi o único marcador cujo valor do parâmetro de Hurst é mais elevado (superior a 0,5) para os TExcel em relação aos demais. Apesar disso, este valor não permite afirmar com segurança a existência de uma estrutura auto-semelhante no texto, segundo esse marcador. Poderíamos reconhecer auto-semelhança se o valor estimado fosse superior a 0,8. Por um lado, esse resultado leva-nos a concluir que este método pode ser melhor explorado com outras configurações para se tentar melhorar os seus resultados em tarefas de avaliação da qualidade e estética do texto. Por outro, a estimativa de apenas um valor acima de 0,5 dos 21 marcadores propostos, pode pôr em causa a eficácia do uso de auto-semelhança para aferir a qualidade de um texto, servindo isto, no entanto, de motivação para investigações futuras. Porém, os resultados obtidos neste trabalho, não permitem validar as desconfianças levantadas em [Cordeiro2015].

Na tentativa de verificar com maior precisão os atributos discriminadores da qualidade e estética do texto, aplicou-se o teste *t-Student* aos resultados obtidos na segunda fase das nossas experiências. Esta fase consistiu na aplicação dos 21 marcadores selecionados aos textos completos. Os resultados obtidos mostraram que 13 dos 21 marcadores forneceram evidência suficiente para diferenciar TExcel e TMeds. Desses marcadores, três são de vocabulário do texto (a taxa de repetição, sofisticação do vocabulário e densidade do vocabulário), três são de coerência baseada em entidades (transições S $\rightarrow$ S, S $\rightarrow$ O e O $\rightarrow$ O), dois são de coesão lexical (similaridade lexical entre parágrafos iniciais e finais e a sobreposição de palavras em frases adjacentes), dois baseados em part of speech (proporção de pronomes por substantivos e diversidade de palavras de conteúdo) e dois de complexidade léxica (comprimento médio da frase e da palavra). Com base nesses dados, temos evidências suficientes para concluir que o nosso método pode ser utilizado para destingir a qualidade e estética de textos.

Finalmente, podemos afirmar que as duas experiências realizadas deram-nos indicadores positivos para tarefas de avaliação da qualidade e estética do texto, com realce nas da segunda etapa. No geral, 13 dos 21 marcadores testados forneceram evidências suficientes para a distinção de TExcel e TMeds. Portanto, esses marcadores podem ser bastante úteis em tarefas de avaliação da qualidade e estética do texto. As aplicações de PLN, especializadas em tarefas semelhantes, podem explorá-las com maior precisão para uma possível validação das mesmas.

5.2 Trabalho Futuro

Durante o desenvolvimento do nosso trabalho deparámo-nos com inúmeras situações de caráter relevante para futuras investigações.
Em primeira instância pretende-se aprimorar a metodologia proposta, sugerindo novas configurações, para testar a eficácia do método de deteção de auto-semelhança nas tarefas de avaliação da qualidade e estética do texto. Essas configurações passam também por testar outros marcadores linguísticos (por exemplo as figuras de estilos) ainda não experimentados.
Outro dado importante para um trabalho futuro é a melhoria dos tempos de processamento. Uma revisão da complexidade dos algoritmos utilizados e uma paralelização do processamento, podem ser bons indicadores para se efetivar essa melhoria.
Finalmente pretende-se construir uma biblioteca que incorpore os 13 marcadores linguísticos discriminadores de qualidade de texto e, disponibiliza-la livremente para a comunidade, a fim de servir de ferramenta de trabalho para pesquisas cientificas relacionadas.

Bibliografia

[Antiqueira2005] Lucas Antiqueira, MGV Nunes, ON Oliveira Jr, and Luciano da F Costa. Modelando textos como redes complexas. In Anais do III Workshop em Tecnologia da Informação e da Linguagem Humana, pages 22–26, 2005. 12, 13

[Antiqueira2007] L. Antiqueira, M. G. V. Nunes, O. N. Oliveira Jr., and L. da F. Costa. Strong correlations between text quality and complex networks features. 373:811–820, 2007. 12, 13

[Barzilay2008] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. Computational Linguistics, 34(1):1–34, 2008. 10, 18, 25

[Beers2009] Scott F. Beers and William E. Nagy. Syntactic complexity as a predictor of adolescent writing quality: Which measures? which genre? 22:185–200, 2009. 11

[Berman2002] Ruth Berman and Ludo Verhoeven. Cross-linguistic perspectives on the development of text-production abilities: Speech and writing. Written Language & Literacy, 5(1):1–43, 2002. 8

[Biber1989] Douglas Biber. A typology of english texts. Linguistics, 27(1):3–44, 1989. 7

[Browne2014] Charles Browne. A new general service list: The better mousetrap we've been looking for. Vocabulary Learning and Instruction, 3(2):1–10, 2014. 39

[Burstein2010] Jill Burstein, Joel R. Tetreault, and Slava Andreyev. Using entity-based features to model coherence in student essays. In Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA, pages 681–684. The Association for Computational Linguistics, 2010. 10

[Carroll1964] John B. Carroll. Language and thought. Prentice-Hall, Englewood Cliffs, New Jersey, 1964. 7

[Cavaco2010] Afonso Miguel Cavaco and Dulce Várzea. Contribuição para o estudo da leitura de folhetos informativos nas farmácias portuguesas. Revista Portuguesa de Saúde Pública, 28(2):179–186, 2010. 5

[Chall1995] Jeanne Sternlicht Chall and Edgar Dale. Readability revisited: The new Dale-Chall readability formula. Brookline Books, 1995. 5

[Charroles1988] M. Charroles. Introdução aos problemas da coerência dos textos: abordagem teórica e estudo das práticas pedagógicas," o texto: leitura e escrita. Technical report, Campinas, SP: Pontes, pp. 39–85„ 1988. 9

[Cheung2010] Jackie Chi Kit Cheung and Gerald Penn. Entity-based local coherence modelling using topological fields. In Jan Hajic, Sandra Carberry, and Stephen Clark, editors, ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, pages 186–195. The Association for Computer Linguistics, 2010. 10

[Cordeiro2015] João Cordeiro, Pedro RM Inácio, and Diogo AB Fernandes. Fractal beauty in text. In Portuguese Conference on Artificial Intelligence, pages 796–802. Springer, 2015. 4, 6, 12, 47, 49, 51

[Covington2010] Michael A Covington and Joe D McFall. Cutting the gordian knot: The moving-average type–token ratio (mattr). Journal of quantitative linguistics, 17(2):94–100, 2010. 8

[Crossley2014] Scott A. Crossley and Danielle S. McNamara. Does writing development equal writing quality? a computational investigation of syntactic complexity in l2 learners. 26:66–79, 2014. 11, 19

[Crossley2016] Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. Behavior research methods, 48(4):1227–1237, 2016. 9, 10, 19, 32, 49

[Crowhurst1983] Marion Crowhurst. Syntactic complexity and writing quality: A review. 8:1, 1983. 11

[Daller2003] Helmut Daller, Roeland Van Hout, and Jeanine Treffers-Daller. Lexical richness in the spontaneous speech of bilinguals. Applied linguistics, 24(2):197–222, 2003. 7

[Dijk1983] Teun Adrianus Dijk and Walter Kintsch. Strategies of discourse comprehension. Academic Press, New York, NY [u.a.], 1983. Literaturverz. S. 387 - 404. 9

[Dubay2004] Willian H Dubay. The principles of readability a brief introduction to readability research. Impact Information, Costa Mesa, CA, 21:489–508, 2004. 5

[Duran2004] Pilar Durán, David Malvern, Brian Richards, and Ngoni Chipere. Developmental trends in lexical diversity. Applied Linguistics, 25(2):220–242, 2004. 8

[Elliot2003] S. Elliot. Intellimetric: From here to validity," automated essay scoring: A cross-disciplinary perspective. page pp. 71–86, 2003. 11

[Ertmer2002] Peggy A Ertmer, Hua Bai, Chaoyan Dong, Mohammed Khalil, Sung Hee Park, and Ling Wang. Online professional development: Building administrators' capacity for technology leadership. Journal of Computing in teacher Education, 19(1):5–11, 2002. 7

[Fergadiotis2013] Gerasimos Fergadiotis, Heather H Wright, and Thomas M West. Measuring lexical diversity in narrative discourse of people with aphasia. American Journal of Speech-Language Pathology, 22(2):S397–S408, 2013. 7

[Fernandes2014] Diogo AB Fernandes, Miguel Neto, Liliana FB Soares, Mário M Freire, and Pedro RM Inácio. A tool for estimating the hurst parameter and for generating self-similar sequences. In Proceedings of the 2014 Summer Simulation Multiconference, page 40. Society for Computer Simulation International, 2014. 40

[Flower1981] Linda Flower and John R Hayes. A cognitive process theory of writing. College composition and communication, 32(4):365–387, 1981. 6

[Freitas2013] Alison RP Freitas and Valéria D Feltrim. Análise automática de coerência usando o modelo grade de entidades para o português. In Proceedings of the IX Brazilian Symposium in Information and Human Language Technology, pages 69–78, 2013. 7, 9, 10, 11, 34

[Freitas2013a] A. R. P. Freitas. Análise automática de coerência usando o modelo grade de entidades para o português. Master's thesis, Universidade Estadual de Maringã, 2013. 9, 10

[Graesser2004] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. Coh-metrix: Analysis of text on cohesion and language. Behavior research methods, instruments, & computers, 36(2):193–202, 2004. 9, 10, 19, 31

[Grosz1995] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. Computational Linguistics, 21(2):203–225, 1995. 10

[Guinaudeau2013] Camille Guinaudeau and Michael Strube. Graph-based local coherence modeling. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers, pages 93–103. The Association for Computer Linguistics, 2013. 10, 18

[Harmon2005] Janis M Harmon, Wanda B Hedrick, and Karen D Wood. Research on vocabulary instruction in the content areas: Implications for struggling readers. Reading & Writing Quarterly, 21(3):261–280, 2005. 6

[Herdan1960] Gustav Herdan. Quantitative linguistics. Butterworth, London, 1960. 7

[Higgins2004] Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. Evaluating multiple aspects of coherence in student essays. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, 2004. 10

[Hunt1970] Kellogg W. Hunt. Syntactic maturity in schoolchildren and adults. 35:iii, 1970. 11

[J.Burstein2003] M. Chodorow J. Burstein and C. Leacock. Criterion: Online essay evaluation: An application for automated evaluation of test-taker essays. Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence, Acapulco, Mexico, 2003. 11

[Jagaiah2017] Thilagha Jagaiah. Analysis of syntactic complexity and its relationship to writing quality in argumentative essays. 2017. 11

[Johansson2009] Victoria Johansson. Lexical diversity and lexical density in speech and writing: a developmental perspective. Working Papers in Linguistics, 53:61–79, 2009. 6, 17, 24, 48

[Johansson2009a] Victoria Johansson. Developmental aspects of text production in writing and speech. Travaux de l'Institut de Linguistique de Lund, 48, 2009. 8

[Koizumi2012a] Rie Koizumi and Yo In'nami. Withdrawn: Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. System, 40(4):522–532, 2012. 8

[Koppel2006] Moshe Koppel, JONATHAN Schler, Shlomo Argamon, and JAMES Pennebaker. Effects of age and gender on blogging. In AAAI 2006 spring symposium on computational approaches to analysing weblogs, pages 1–7, 2006. 43

[Landauer2003] Thomas K Landauer. Automated scoring and annotation of essays with the intelligent essay assessor. Automated essay scoring: A crossdisciplinary perspective, 2003. 10

[Lapata2005] Mirella Lapata and Regina Barzilay. Automatic evaluation of text coherence: Models and representations. In IJCAI, volume 5, pages 1085–1090, 2005. 9

[Laufer2013] Batia Laufer. Vocabulary and writing. The Encyclopedia of Applied Linguistics, 2013. 6

[Louis2012] Annie Louis. Automatic metrics for genre-specific text quality. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 54–59. Association for Computational Linguistics, 2012. 5, 13, 14

[Louis2013] Annie Priyadarshini Louis. Predicting text quality: metrics for content, organization and reader interest. PhD thesis, University of Pennsylvania, 2013. 6

[Lu2011] Xiaofei Lu. A corpus-based evaluation of syntactic complexity measures as indices of college-level esl writers' language development. 45:36–62, 2011. 11

[Machado2012] Áurea Maria Bezerra Machado and Márcio Luiz Corrêa Vilaça. A importância da coesão e da coerência em nossos textos. Cadernos do CNLF (CiFEFil), v. XVI, 16(04):76–83, 2012. 8, 9, 18

[Malvern2004] David Malvern, Brian J Richards, Ngoni Chipere, and P Purán. Lexical diversity and language development. Springer, 2004. 7, 8

[Manning2014] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pages 55–60, 2014. xiii, 38

[Martins2016] Mário Martins. A diversidade lexical na escrita de textos escolares. Fórum Linguístico, 13(1):1068–1082, 2016. 7, 8

[McCarthy2005] Philip M McCarthy. An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (mtld). Dissertation Abstracts International, 66:12, 2005. 8, 17

[McCarthy2007] Philip M McCarthy and Scott Jarvis. vocd: A theoretical and empirical evaluation. Language Testing, 24(4):459–488, 2007. 8

[McCarthy2010] Philip M McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. Behavior research methods, 42(2):381–392, 2010. 6, 7, 8, 20, 49

[McNamara2010] Danielle S. McNamara, Scott A. Crossley, and Philip M. McCarthy. Linguistic features of writing quality. 27:57–86, 2010. 14

[McNamara2011] Danielle S McNamara, Arthur C Graesser, Zhiqiang Cai, and Jonna M Kulikowich. Coh-metrix easability components: Aligning text difficulty with theories of text comprehension. In annual meeting of the American Educational Research Association, New Orleans, LA, 2011. 10

[Mesgar2014] Mohsen Mesgar and Michael Strube. Normalized entity graph for computing local coherence. In V. G. Vinod Vydiswaran, Amarnag Subramanya, Gabor Melli, and Irina Matveeva, editors, Proceedings of TextGraphs@EMNLP 2014: the 9th Workshop on Graph-based Methods for Natural Language Processing, October 29, 2014, Doha, Qatar, pages 1–5. The Association for Computer Linguistics, 2014. 10, 18, 25, 49

[Oestling2017] Robert Östling and Gintare Grigonyte. Transparent text quality assessment with convolutional neural networks, 2017. 6, 12, 13

[Olinghouse2009] Natalie G Olinghouse and Jacqueline T Leaird. The relationship between measures of vocabulary and narrative writing quality in second-and fourth-grade students. Reading and Writing, 22(5):545–565, 2009. 6, 48

[Olinghouse2013] Natalie G. Olinghouse and Joshua Wilson. The relationship between vocabulary and writing quality in three genres. Reading and Writing, 26(1):45, January 2013. 6, 17, 23

[Ortega2003] L. Ortega. Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing. 24:492–518, 2003. 11

[Pierre1960] Guiraud Pierre. Problèmes et méthodes de la statistique linguistique, 1960. 7

[Pitler2008] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In Proceedings of the conference on empirical methods in natural language processing, pages 186–195. Association for Computational Linguistics, 2008. 5, 6, 10, 13, 19, 29, 32, 48

[Richards1987] Brian Richards. Type/token ratios: What do they really tell us? Journal of child language, 14(2):201–209, 1987. 7

[Richards1997] Brian J Richards and David D Malvern. Quantifying lexical diversity in the study of language development. University of Reading, Faculty of Education and Community Studies, 1997. 7, 8

[Scherer2002] Sabrina Scherer, F Casarim, Patrícia Zart, and A Ramos. Perfil evolutivo da relação type/token de crianças de 3 a 5 anos de idade. Porto Alegre: Trabalho de especialização, CEFAC, 2002. 7

[Schriver1989] Karen A Schriver. Evaluating text quality: The continuum from text-focused to reader-focused methods. IEEE Transactions on professional communication, 32(4):238–255, 1989. 5

[Silva2015] Leandro Lago da Silva and Valéria Delisandra Feltrim. Análise automática de coerência textual em resumos científicos: Avaliando quebras de linearidade (automatic analysis of textual coherence in scientific abstracts: Evaluating linearity breaks). In Cláudia Freitas and Alexandre Rademaker, editors, Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology, STIL 2015, Natal, Brazil, November 4-7, 2015, pages 45–49. Sociedade Brasileira de Computação, 2015. 10

[Souza2013] Vinícius Mourão Alves de Souza and Valéria Delisandra Feltrim. A coherence analysis module for scipo: providing suggestions for scientific abstracts written in portuguese. J. Braz. Comp. Soc., 19(1):59–73, 2013. 10, 11

[Stewart1979] Murray F Stewart and Cary H Grobe. Syntactic maturity, mechanics of writing, and teachers' quality ratings. Research in the Teaching of English, 13(3):207–215, 1979. 11

[Stroemqvist2002] Sven Strömqvist, Victoria Johansson, Sarah Kriz, Hrafnhildur Ragnarsdottir, Ravid Aisenman, and Dorit Ravid. Toward a cross-linguistic comparison of lexical quanta in speech and writing. Written Language & Literacy, 5(1):45–67, 2002. 8

[Templin1957a] Mildred Templin. Certain language skills in children: Their development and interrelationships (monograph series no. 26). Minneapolis: University of Minnesota, The Institute of Child Welfare, 1957. 7

[Thomas2005] Dax Thomas. Type-token ratios in one teacher's classroom talk: An investigation of lexical complexity. United Kingdon: University of Birmingham, 2005. 7

[Todirascu2013] Amalia Todirascu, Thomas François, Nuria Gala, Cédric Fairon, Anne-Laure Ligozat, and Delphine Bernhard. Coherence and cohesion for the assessment of text readability. Natural Language Processing and Cognitive Science, 11:11–19, 2013. 9, 10, 18, 19, 29, 32, 48, 49

[Toernqvist2015] Jennifer Törnqvist. Using coh-metrix to investigate changes in student texts: Comparing student writing from 1999 and 2009, 2015. 10

[Treffers-Daller2013] Jeanine Treffers-Daller. Measuring lexical diversity among L2 learners of French. John Benjamins Amsterdam/Philadelphia, 2013. 8

[VanDijk1980] Teun Adrianus Van Dijk. Text and context explorations in the semantics and pragmatics of discourse. 1980. 9

[Verheijen2016] Lieke Verheijen. Linguistic characteristics of dutch computer-mediated communication: Cmc and school writing compared. 2016. 8

[Wu1993] Trong Wu. An accurate computation of the hypergeometric distribution function. ACM Transactions on Mathematical Software (TOMS), 19(1):33–43, 1993. 8

[Youmans1990] Gilbert Youmans. Measuring lexical style and competence: The type-token vocabulary curve. Style, pages 584–599, 1990. 7