UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

GABRIELLE PIEZZOTI OLIVEIRA

ESTUDO COMPARATIVO DA MORFOLOGIA TEXTUAL DE LAUDOS MÉDICOS E DE TEXTOS DE PROPÓSITO GERAL

MONOGRAFIA

CAMPO MOURÃO
2017

GABRIELLE PIEZZOTI OLIVEIRA

ESTUDO COMPARATIVO DA MORFOLOGIA TEXTUAL DE LAUDOS MÉDICOS E DE TEXTOS DE PROPÓSITO GERAL

Trabalho de Conclusão de Curso de graduação apresentado à disciplina de Trabalho de Conclusão de Curso 2, do Curso de Bacharelado em Ciência da Computação do Departamento Acadêmico de Computação da Universidade Tecnológica Federal do Paraná, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Lucio Geronimo Valentin

CAMPO MOURÃO
2017

ATA DA DEFESA DO PROJETO DE TCC

Às **14:30** do dia **21 de junho de 2017** foi realizada na sala **E104** da UTFPR-CM a sessão pública da defesa do projeto de Trabalho de Conclusão do Curso de Bacharelado em Ciência da Computação do(a) acadêmico(a) **Gabrielle Piezzoti Oliveira**. Estavam presentes, além do(a) acadêmico(a), os membros da banca examinadora composta por: **Prof. Dr. Lucio Geronimo Valentin** (orientador), **Prof. Dr. Rodrigo Campiolo** e **Prof. Dr. Frank Helbert Borsato**. Inicialmente, o(a) acadêmico(a) fez a apresentação do seu trabalho, sendo, em seguida, arguido(a) pela banca examinadora. Após as arguições, sem a presença do(a) acadêmico(a), a banca examinadora o(a) considerou _____ na disciplina de Trabalho de Conclusão de Curso **2** e atribuiu, em consenso, a nota _____ (_____). Este resultado foi comunicado ao(à) acadêmico(a) e aos presentes na sessão pública e, posteriormente, deverá ser registrado no sistema acadêmico pelo professor responsável de TCC. Em seguida foi encerrada a sessão e, para constar, foi lavrada a presente Ata que segue assinada pelos membros da banca examinadora, após lida e considerada conforme.

Observações: _____

_____

_____

_____

_____

_____

Campo Mourão, **21 de junho de 2017**

Prof. Dr. Rodrigo Campiolo
Membro 1

Prof. Dr. Frank Helbert Borsato
Membro 2

Prof. Dr. Lucio Geronimo Valentin
Orientador

Resumo

Oliveira, G. P.. Estudo Comparativo da Morfologia Textual de Laudos Médicos e de Textos de Propósito Geral. 2017. 58. f. Monografia (Curso de Bacharelado em Ciência da Computação), Universidade Tecnológica Federal do Paraná. Campo Mourão, 2017.

Contexto: O entendimento da linguagem humana por parte de computadores já é questão de pesquisa há anos e pode ser usada com diferentes finalidades. Esta monografia se utiliza do Processamento de Linguagem Natural (PLN) para realizar um levantamento morfológico do conteúdo de laudos médicos. Tais informações podem auxiliar pesquisas futuras desse tipo de texto, pois há uma carência na literatura de estudos sobre sua estrutura.
Objetivo: O objetivo é identificar as características morfológicas predominantes nesse tipo de texto e analisar como diferem de outros tipos de textos. Para isso, são usados dois *corpora*: um de cunho jornalístico e um de cunho acadêmico.
Método: O primeiro passo da pesquisa é a formatação de tais *corpora*, deixando-os em um formato compatível com a ferramenta de análise morfológica usada, seguido por uma limpeza, para correção e remoção de elementos indesejados. Os laudos exigiram ainda etapas extras, para a seleção das amostras que compõem o *corpus*. Segue-se então com o processamento dos *corpora* e o levantamento das seguintes informações: classes gramaticais, lemas, unigramas, bigramas e trigramas mais frequentes.
Resultados: As estatísticas levantadas em cima dos dados obtidos no processamento mostram uma variação considerável entre o conteúdo morfológico dos laudos e dos demais *corpora*, reforçando as hipóteses empíricas de que há uma limitação e especificidade nos laudos, além de uma diferença significativa na proporção de adjetivos, verbos e números, para com os outros textos.
Conclusões: A análise traçou o perfil morfológico dos laudos como sendo um texto com mais substantivos, números e adjetivos que os demais, mas com menos verbos, pronomes e determinantes. Além disso, possui frases mais simples e diretas, com um vocabulário limitado. Outras pesquisas podem ser feitas, variando os *corpora*, para enriquecer a análise morfológica apresentada nesta monografia.
Palavras-chaves: Processamento de Linguagem Natural. Morfologia. Laudos Médicos. Linguística de Corpus. Etiquetagem Morfológica.

Abstract

Oliveira, G. P.. Comparative Study of the Textual Morphology of Medical Reports and General Purpose Texts. 2017. 58. f. Monograph (Undergradute Program in Computer Science), Federal University of Technology – Paraná. Campo Mourão, PR, Brazil, 2017.

Context: The understanding of human speech by computers has been a research question for years and can be applied to different goals. This monograph uses Natural Language Processing (NLP) to perform a morphological survey of the content of medical reports. This data can help future studies of this type of text, since there is a lack of analysis about medical reports in the literature.
Objective: The goal is to identify the predominant features of this type of text and analyze how it differs from other types of texts. In order to do this, an academic and a journalistic *corpora* will be used.
Method: The first step was the formatting of these *corpora*, so they are compatible with the analysis tool, followed by an cleanup for correction and removal of unwanted elements. The medical reports required extra steps, for the selection of the 500 reports that compose the *corpus*. The next thing is the processing of the *corpora* and the gathering of the following informations: the most frequent parts of speech, lemmas, unigrams, bigrams and trigrams.
Results: The statistics extracted from the data provided by the tool show a considerable amount of variation between the morphological content of the medical reports and the other texts, which reinforces the empirical hypotheses: there is a significant difference in the specification of the medical reports and the proportion of adjectives, verbs and numbers, compared to the other types of texts.
Conclusion: The analysis described the morphological profile of the medical reports as being a type of text with more nouns, numbers and adjectives than the others, but with fewer verbs, pronouns and determinants. Moreover, they have simpler and more direct sentences, with a limited vocabulary. Further research could be done, varying the *corpora*, to enrich the morphological analysis presented in this monograph.
Keywords: Natural Language Processsing. Morphology. Medical reports. Corpus Linguistics. POS-Tagging

Lista de figuras

Lista de tabelas

Sumário

Capítulo 1

Introdução

A comunicação com os computadores costuma se dar por meio de Linguagens Formais, compostas por instruções precisas, de estrutura bem delimitada e sem ambiguidade. A linguagem que os humanos usam para se comunicar entre si, por outro lado, nem sempre é clara, apresenta muitas ambiguidades, usa gírias e depende do contexto.

O processamento da linguagem humana por parte de computadores é chamado de Processamento de Linguagem Natural (PLN). De forma geral, busca-se fazer com que um computador seja capaz de entender um discurso humano em sua forma natural. Pesquisas envolvendo PLN cobrem as mais diversas áreas e têm as mais variadas finalidades, como análise de sentimentos, tradução automática e geração de resumos.

Uma das etapas do PLN é a análise morfológica. Ela gera um texto anotado, no qual é associada uma etiqueta a cada palavra, indicando sua classe gramatical. Esta monografia usa dessa análise para realizar o levantamento morfológico do conteúdo de laudos médicos. Por "laudo médico", entende-se o documento criado por um médico especialista onde este descreve os elementos observados durante um exame médico.

Assim, esta monografia busca responder a seguinte questão de pesquisa: Quais são as características morfológicas predominantes em textos de laudos e como diferem de outros tipos de texto?

O objetivo é levantar estatísticas a respeito dos laudos, de modo a se criar um perfil morfológico desse tipo de texto. Para isso, são consideradas as classes gramaticais mais frequentes, e ainda as palavras, lemas, bigramas e trigramas mais frequentes. Um *corpora* acadêmico e um jornalístico são usados como base de comparação.

Os resultados obtidos podem auxiliar futuros estudos desse tipo de texto, pois há uma carência de pesquisas científicas com tal foco, que apresentem uma análise do formato e da estrutura morfológica de laudos, não apenas observações empíricas.

O primeiro passo da pesquisa foi o levantamento teórico a respeito da área, e uma

apuração dos *corpora* e ferramentas de etiquetagem disponíveis para português do Brasil. A seguir, os *corpora* foram obtidos (os laudos usados vêm de uma base privada), passaram por um pré-processamento (formatação e limpeza) e só então foram processados para extração de informações. Por fim, tais dados foram analisados e as diferenças para com os demais tipos de textos foram levantadas. O processamento foi feito por meio de um portal WEB chamado MorfoX, desenvolvido em um projeto paralelo. O portal usa HTML5 e CSS3, tem o Freeling como ferramenta morfológica, e todas as funcionalidades são em JavaScript.

Esta monografia estrutura-se em cinco capítulos, incluindo este. No Capítulo 2, apresenta-se a fundamentação teórica que embasa o trabalho, que conta com a definição de conceitos de PLN, técnicas de etiquetagem e exemplos de etiquetadores e *corpora*. O Capítulo 3 apresenta toda a metodologia adotada no trabalho, com detalhamento do processo de seleção, preparação e processamento dos *corpora*. O Capítulo 4 traz os resultados obtidos, as estatísticas extraídas desses dados e uma análise dessas informações, respondendo a questão de pesquisa. Por fim, o Capítulo 5 apresenta uma conclusão sobre tais resultados e suas implicações.

Capítulo 2

Referencial Teórico

Com computadores ocupando cada vez mais espaço em nosso dia-a-dia, no aspecto pessoal, social e profissional, é necessário que eles entendam o que queremos dizer a eles (direta ou indiretamente). Para dar instruções aos computadores sobre tarefas a serem executadas, usa-se as linguagens de programação. Para mediar e facilitar a interação das pessoas com o computador, tem-se interfaces gráficas. Mas fazer um computador entender a linguagem tal como as pessoas a usam para se comunicar, campo denominado Processamento de Linguagem Natural (PLN), tem sido um desafio desde a década de 50 e agrega pesquisadores de diversas áreas, sobretudo da Linguística e da Inteligência Artificial (SILVA et al., 2007).

O PLN trabalha com conteúdos digitais (áudio/texto) que contenham alguma forma de discurso humano em linguagem natural. Antes de seu uso, esses conteúdos passam por um pré-processamento no qual são limpos (remoção de elementos não linguísticos ou ruídos) e segmentados (separação de sentenças e/ou palavras). Cada elemento é analisado pelo computador e processado de acordo com o objetivo, que pode ser, dentre outros, extração de informação, geração de resumos, tradução automática, correção ortográfica, análise de sentimentos, reconhecimento de fala e clusterização (VIEIRA; LIMA, 2001).

2.1. Corpora

Como uma das principais fontes de dados para o Processamento de Linguagem Natural são textos, *corpora* são vitais para pesquisas na área. Segundo Wynne (2005, p. 23, tradução nossa), "um *corpus* é uma coleção de partes de textos em forma eletrônica, selecionados de acordo com critérios externos para representar, tanto quanto o possível, uma linguagem ou variedade linguística como fonte de dados para pesquisas linguísticas"[1].

[1] Original: A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.

Assim, um *corpus* linguístico é um conjunto de textos autênticos (realmente produzidos por falantes da língua em questão) que compõem uma amostra representativa da variação linguística que deseja-se estudar. Propriedades como tamanho do *corpus*, estrutura, amostragem e direitos autorais devem ser consideradas quando trabalha-se com *corpora*.

A exemplo de *corpora* em português, pode-se citar o NILC/São Carlos (assim batizado pelo Linguateca[2] (SANTOS, 2009) em 1999), que contém 32 milhões de palavras. Esse *corpus* foi criado para auxiliar o projeto ReGra (1993) (PARDO, 2000), um revisor gramatical para português, parceria do Itautec e do Núcleo Interinstitucional de Linguística Computacional (NILC)[3] da Universidade de São Paulo (USP), *campus* de São Carlos. O *corpus* é formado por textos corrigidos (material jurídico, jornalístico, didático, literário e técnico científico), semicorrigidos (material universitário, como artigos, relatórios e teses) e não-corrigidos (redações de vestibular) (PINHEIRO; ALUÍSIO, 2003). Pode parecer abrangente, mas as áreas não são representadas de forma homogênea no *corpus* (há apenas alguns textos científicos, enquanto que textos jornalísticos representam mais de 70% do *corpus*), e há problemas na formatação das amostras, como irregularidades e a junção de diferentes textos em um mesmo arquivo.

Em virtude dos textos jornalísticos serem maioria no *corpus* NILC/São Carlos, em 2002, o projeto Processamento Computacional do Português (Portugal) (SANTOS, 2000) criou um *corpus* jornalístico baseado nele, chamado CETENFolha (Corpus de Extractos de Textos Electrónicos NILC/Folha de São Paulo) (LINGUATECA, 2016). Esse *corpus* é formado por extratos de todas as edições do jornal Folha de São Paulo de 1994 e conta com quase 24 milhões de palavras. Sua limitação se dá no âmbito temporal, por englobar amostras de apenas um ano. Ambos os *corpus* são disponibilizados online pelo projeto AC/DC (Acesso a corpos/Disponibilização de corpos)[4].

O CETENFolha está incluso na Coleção CHAVE[5], também do Linguateca, que possui ainda os textos completos do jornal Folha de São Paulo de 1994 e 1995, o *corpus* CETENPublico[6] (Corpus de Extractos de Textos Electrónicos MCT/Público), formado por artigos do jornal português Público, e a versão anotada desses *corpora*.

Outro *corpus* de caráter especifico é o CorpusDT (FELTRIM et al., 2001), criado para auxiliar o projeto SciPo (ANTIQUEIRA et al., 2003), um sistema de auxílio à escrita de resumos acadêmicos, também do NILC. O *corpus* é formado por introduções e resumos de dissertações de mestrado e teses de doutorado de várias áreas da Ciência da Computação. São 52 trabalhos, sendo 49 dissertações e 3 teses, datando de 1994 a 2001.

[2] Página do Linguateca: http://www.linguateca.pt/
[3] Página do NILC: http://www.nilc.icmc.usp.br/
[4] Acesso ao AC/DC: http://www.linguateca.pt/ACDC/
[5] Acesso à CHAVE: http://www.linguateca.pt/CHAVE/
[6] Mais informações do CETENPublico: http://www.linguateca.pt/CETEMPublico/informacoes.html

2.2. Morfologia

Tendo-se a base textual, o PLN pode incluir etapas de análises fonética, que lida com a propriedade sonora da palavra, morfológica, que lida com a estrutura da palavra, sintática, que trata a função das palavras dentro de uma sentença, e pragmática-discursiva, que lida com informações de contexto (SILVA et al., 2007). Esta monografia usa da etapa morfológica.

Morfologia é a parte da gramática que estuda a estrutura das palavras e sua classificação em diferentes categorias (classes gramaticais ou *parts of speech*/POS). A respeito da estrutura, as palavras são formadas por unidades significativas chamadas morfemas, são elas: radical, vogal temática, tema, desinência, afixos e vogais/consoantes de ligação. A palavra "carro" é formada por um radical (carr) e uma vogal temática (o), enquanto que "lealdade" é formada por um radical (leal) e um sufixo (-dade). Quanto às classes gramaticais, a língua portuguesa possui 10 categorias elementares: adjetivo, advérbio, artigo, conjunção, numeral, interjeição, preposição, pronome, substantivo e verbo. As palavras de uma mesma categoria compartilham propriedades em comum, como a flexão em número, gênero e grau. (VIEIRA; LIMA, 2001).

Além da estrutura morfológica, uma palavra pode ser representada por seu lema, a unidade elementar da qual se derivam as palavras. No geral, usa-se o infinitivo para representar as variações verbais e o masculino singular para substantivos e adjetivos. Por exemplo, "é", "seremos" e "era" são representados pelo lema "ser", enquanto que "cachorra", "cachorros" e "cachorrinho" são representados por "cachorro" (LUCCA; NUNES, 2002).

2.3. N-Gramas

Agrupamentos de itens em sequência dentro de uma frase (palavras, letras...) são conhecidos como N-gramas, onde N indica o número de elementos. Alguns conjuntos recebem nomes específicos: 1-grama (um item) são chamados de unigramas, enquanto que 2-gramas (dois itens) são chamados de bigramas e 3-gramas (3 itens) são trigramas.

Em PLN, costuma-se desconsiderar pontuação, e pode-se ignorar ou não acentos e *stop words* (palavras consideradas irrelevantes, como preposições e artigos).

Um exemplo de uso da técnica é a predição de texto, no qual calcula-se a probabilidade da palavra seguinte considerando as $N-1$ palavras anteriores e suas características. Modelos de etiquetagem morfológica usam de N-gramas para predizer as classes gramaticais das palavras (JURAFSKY; MARTIN, 2016a).

2.4. Etiquetagem

Para auxiliar diferentes fases do processamento, informações linguísticas são adicionadas ao texto, técnica que recebe o nome de "anotação". A anotação pode ser a nível morfológico,

sintático ou semântico.

A anotação de um *corpus* quanto à classe gramatical das palavras é chamada de etiquetagem morfossintática (*POS-tagging*, em inglês). A cada palavra presente em uma sentença, adiciona-se um marcador (uma etiqueta ou *tag*) indicando a qual classe ela pertence. O computador não sabe o que é um verbo, adjetivo ou substantivo, mas com a anotação morfológica ele pode fazer análises em cima dessas classes e das palavras pertencentes a elas.

Um etiquetador, ou *tagger*, é a ferramenta computacional que realiza essa etiquetagem de forma automática (OTHERO; AYRES, 2014). A entrada é um *corpus* e o conjunto de possíveis etiquetas (*tagset*) e a saída é um *corpus* anotado (*POS-tagged*). O resultado da etiquetagem da frase "O dia parecia não ter fim" seria "O_ART dia_SUB parecia_VRB não_ADVNEG ter_VRB fim_SUB". As etiquetas presentes na sentença são ART artigo, SUB substantivo, VRB verbo e ADVNEG advérbio de negação.

O tamanho do *tagset* pode variar segundo a finalidade do *corpus*. Saber que uma palavra é um verbo pode ser o suficiente para um estudo mas, para outro, pode ser necessário, por exemplo, saber em qual tempo ele está conjugado, o que muda as etiquetas. O *corpus* NILC/São Carlos (ALUÍSIO; AIRES, 2000) usa 37 etiquetas, enquanto que o Penn Treebank (MARCUS et al., 1993) tem 45 e o pioneiro Brown *corpus* (FRANCIS; KUCERA, 1964) tem 87.

Enquanto que alguns *corpora* possuem apenas o texto em si, muitos *corpora* são disponibilizados já anotados morfossintaticamente (o CETENFolha oferece as duas versões). É o caso do Floresta Sintá(c)tica[7], um *corpus* em português anotado pelo PALAVRAS (etiquetador apresentado na seção seguinte), formado por cerca de 6,7 milhões de palavras, de fontes brasileiras e portuguesas. Ele é composto por quatro partes, variando o gênero dos textos, o modo (escrito ou falado) e o grau de revisão manual da etiquetagem, como descrito na página do projeto. O principal sub-corpus é o Bosque, totalmente revisado por linguistas.

Outro exemplo é o Mac-Morpho (ALUÍSIO et al., 2003), hoje em sua terceira versão, o maior *corpus* anotado em português revisado manualmente, disponível online. Ele também usa o PALAVRAS e é composto apenas por textos do português brasileiro, extraídos do jornal Folha de São Paulo. Também foi desenvolvido pelo NILC.

2.5. Etiquetadores

Há dois tipos fundamentais de algoritmos etiquetadores: baseado em regras e probabilístico (aprendizado de máquina). No sistema baseado em regras, um léxico define as possíveis classes para uma palavra e uma base de regras, geralmente, construídas manualmente, é usada para definir qual a classe correta para a palavra. Um exemplo de regra é

[7] Página do Floresta Sintá(c)tica: http://www.linguateca.pt/floresta/corpus.html

<artigo><adjetivo><substantivo>, que diz que uma palavra antecedida por um par artigo-adjetivo é um substantivo.

Quanto ao modelo probabilístico, costuma-se usar Modelos Ocultos de Markov (HMM) ou Entropia Máxima (ME). O objetivo do HMM é encontrar a sequência de etiquetas mais provável para uma sequência de palavras. Dado um *corpus* de treinamento anotado manualmente, as probabilidades são estimadas com base na frequência na qual uma etiqueta é atribuída a uma palavra. Este modelo, no qual os estados são as possíveis etiquetas, considera dois elementos: a probabilidade de haver uma transição do estado T_i para o estado T_j (frequência da primeira etiqueta ser seguida pela segunda) e a probabilidade de um estado emitir uma determinada palavra (frequência com a qual a palavra é associada à etiqueta daquele estado) (JURAFSKY; MARTIN, 2016b). Esta breve descrição corresponde ao modelo bigrama, que considera apenas a etiqueta anterior para calcular a probabilidade da atual, mas costuma-se usar um modelo trigrama, que usa as duas etiquetas anteriores.

No método de ME, a entropia se relaciona à incerteza probabilística. Não se faz nenhuma suposição sobre o *corpus*, as chances são iguais para todas as possibilidades, então a entropia é máxima. Por exemplo, "a" pode ser preposição, artigo ou pronome oblíquo, então a probabilidade inicial de cada seria $\frac{1}{3}$. O modelo é definido como $H \times T$, onde H é o conjunto de contextos (histórias) e T é o conjunto de etiquetas. Dado um *corpus* anotado de treinamento, o modelo define h_i como sendo o contexto disponível ao se predizer t_i, onde $t \in T$ e $h \in H$. O terceiro elemento do modelo são as *features* $f_i(h, t)$. Essas *features* são binárias (1 ou 0), indicando presença ou ausência de determinada informação de contexto. Exemplificando: dada uma história que considera as duas palavras anteriores e suas etiquetas e as duas palavras seguintes, uma *feature* pode checar se a palavra atual (w_i) termina em "-ismo" ou se a etiqueta da palavra anterior (t_{i-1}) é "verbo". Associa-se um parâmetro para cada *feature* para calcular a probabilidade $p(h, t)$. O modelo em detalhes é descrito em Indurkhya e Damerau (2010).

Existem ainda os etiquetadores híbridos, que combinam ambas as técnicas, como o proposto por Brill (1995), que usa aprendizado baseado em transformações (TBL, *transformation based learning*). Tal técnica infere automaticamente um conjunto ordenado de regras, dado um *corpus* de treinamento. As palavras recebem, inicialmente, a etiqueta mais provável. As etiquetas são então trocadas segundo as regras e conta-se os erros de etiquetagem. A transformação que gerar a maior redução de erros é escolhida e as etiquetas correspondentes são aplicadas.

2.5.1. Dificuldades

Por lidar com linguagem natural, um dos grandes desafios da etiquetagem são as ambiguidades (palavras que podem pertencer a mais de uma classe gramatical, como "planta", que pode ser um verbo ou substantivo). Os algoritmos resolvem esse problema olhando para as palavras

vizinhas e verificando nas regras ou nas probabilidades se uma sequência de etiquetas é plausível. Outra solução é combinar etiquetadores, seja um etiquetador colocando as possíveis etiquetas e um segundo resolvendo as ambiguidades, ou combinando-se os resultados deles e escolhendo a etiqueta com base em alguma estratégia de voto (AIRES et al., 2000).

O segundo desafio é a presença de palavras desconhecidas, que não aparecem no *corpus* de treinamento ou no léxico (dicionário). Nesse caso, pode-se usar informações de contexto (palavras vizinhas), mas a abordagem mais comum é analisar a estrutura morfológica da palavra: prefixo, sufixo, presença de maiúscula ou de um caractere específico (como hífen). Por exemplo, uma palavra terminada em "-ismo", provavelmente, será um substantivo.

2.5.2. Ferramentas

Os principais etiquetadores para o português são:

- Aelius[8]: Este etiquetador híbrido brasileiro (arquitetura RAUBT) é um pacote em Python que usa a biblioteca Natural Language Toolkit (NLTK)[9] e está disponível gratuitamente (ALENCAR, 2010).
- PALAVRAS: Eckhard Bick desenvolveu este analisador morfológico e sintático em um projeto de doutorado. A ferramenta usa o paradigma de *Constraint Grammar* e faz parte do projeto VISL (Visual Interactive Language Learning)[10] (BICK, 2000).
- MXPOST[11]: Este etiquetador JAVA, da Universidade de Edimburgo, usa Entropia Máxima para predizer as classes gramaticais. Foi treinado com o Wall Street Journal *corpus*. O modelo para português foi provido pelo NILC-USP. (RATNAPARKHI, 1996)
- TreeTagger[12]: Criado na Universidade de Stuttgart, usa árvores de decisão binárias e, além da etiquetagem, faz a lematização do texto. A probabilidade do 3-grama é determinada pelo caminho percorrido na árvore até uma folha (HELMUT, 1994).
- FreeLing[13]: Mantido pela Universidade Politécnica da Catalunha, é um conjunto de bibliotecas multilíngue feito em C++ e disponível sob a GNU AGPL. A base do etiquetador é um HMM 3-grama. O modelo em português foi feito pelo grupo ProLNAt[14], da Universidade de Santiago de Compostela (PADRÓ et al., 2010). É o etiquetador oficial do Sketch Engine[15], um sistema de gerenciamento e análise de *corpus*.

[8] Acesso ao Aelius: http://aelius.sourceforge.net/
[9] Acesso ao NLTK: http://www.nltk.org/
[10] Acesso ao VISL: http://visl.sdu.dk/
[11] Acesso ao MXPOST: http://www.inf.ed.ac.uk/resources/nlp/local_doc/MXPOST.html
[12] Acesso ao TreeTagger: http://www.cis.uni-muenchen.de/∼schmid/tools/TreeTagger/
[13] Acesso ao Freeling: http://nlp.lsi.upc.edu/freeling/
[14] Modelo em português do ProLNAt: http://gramatica.usc.es/pln/tools/freeling/download.htmlx
[15] Acesso ao Sketch Engine: https://www.sketchengine.co.uk/

Capítulo 3

Métodos

Para se realizar estudos de laudos médicos é necessário conhecer a forma e a estrutura desse tipo de texto. Especificidades desses textos, como a predominância de adjetivos e substantivos e ausência de orações complexas e verbos, podem ser vistas empiricamente, como nos exemplos: "Útero com morfologia homogênea" (o aspecto visual do útero está homogêneo) e "Tecido subcutâneo sem alterações" (não há alterações nos tecidos). Porém, não encontra-se na literatura estudos que façam essa análise cientificamente.

Assim, a questão de pesquisa que orienta esta monografia é: Quais são as características morfológicas predominantes em textos de laudos médicos e como diferem de outros tipos de textos?

Para respondê-la, as seguintes atividades foram realizadas:

1. Seleção dos *corpora* para comparação;
2. Preparação dos *corpora*;
3. Processamento dos *corpora*;
4. Análise dos resultados.

3.1. Seleção dos *Corpora*

A comparação dos laudos médicos a apenas um tipo de texto poderia criar dúvidas quanto ao viés dos resultados obtidos, então decidiu-se usar dois tipos de textos no estudo, de diferentes áreas.

A escolha de tais *corpora* se deu pela organização de seus textos, por não estarem anotados e por serem *corpora* em português do Brasil que se encontram disponíveis publicamente. Muitos *corpora* são privados, pois são elaborados com uma finalidade específica, partindo de uma intenção de pesquisa, como é o caso do *corpus* de laudos utilizado nesta monografia.

Assim, a princípio, escolheu-se o CorpusDT, de cunho acadêmico (Computação) e o CETENFolha, de cunho jornalístico. Posteriormente, o CETENFolha foi substituído por um *corpus* similar, também formado por artigos das edições do Folha de São Paulo de 1994.

3.1.1. CorpusDT

O CorpusDT conta com 100 arquivos, num total de 52 textos acadêmicos. Alguns trabalhos estão em arquivos únicos, enquanto que outros foram separados em até 5 partes, de acordo com o tamanho. Não há qualquer conteúdo adicional nos arquivos, somente o texto.

As áreas de pesquisa dentro do *corpus* estão distribuídas conforme a Tabela 3.1.

Tabela 3.1. CorpusDT: número de trabalhos em cada área

Área	Dissertação	Tese
Banco de dados	3	0
Inteligência Computacional	7	0
Engenharia de Software	15	1
Hipermídia	12	0
Sistemas Digitais	1	0
Sistemas Distribuídos e Programação Concorrente	9	2
Computação Gráfica e Processamento de Imagens	1	0
Computação de Alto Desempenho	1	0
Total	49	3

Fonte: Extraído de Feltrim et al. (2001)

3.1.2. CETENFolha

O CETENFolha, por sua vez, é formado por um arquivo único, com 340.947 extratos do jornal. Os extratos são, em geral, dois parágrafos de um artigo, e têm a forma:

```
<ext id= cad="" sec="" sem="">
<s> <t>  </t> </s>
<s> <a>  </a> </s>
<p>
<s> </s>
</p>
</ext>
```

Figura 3.1. Formato dos extratos no CETENFolha
Fonte: Elaborado pela autora.

A etiqueta **<ext>** delimita um extrato, identificado por um número e pelo caderno, seção e semestre de publicação; **<t>** indica o título e **<a>** o autor, enquanto que **<p>**

corresponde a um parágrafo e **<s>** a uma frase.

No entanto, em uma análise inicial do *corpus*, notou-se problemas em alguns extratos, como frases incompletas ou separadas incorretamente. A Figura 3.2 mostra alguns exemplos.

```
<s> julho: 15 °C; </s>

<s> ... </s>

<s> Em "M.E.D.I.T." </s>
<s> , o artista se apresenta numa série de sete fotos registrando
o progressivo emaranhamento de seu rosto por um fio . </s>

<s> "Ele veio aqui e liberou US$ 5 milhões, 1% do valor da obra." </s>
<s> , declarou . </s>
```

Figura 3.2. Problemas no CETENFolha

Fonte: Elaborado pela autora.

Então, optou-se por não usar o *corpus* tal como é dado, mas sim fazer uma nova extração de artigos do Folha de São Paulo, por meio da coleção CHAVE.

3.1.2.1. O Novo *Corpus*

Para ter acesso à coleção CHAVE, na Linguateca, é necessário solicitar uma senha, na página da coleção. O acesso foi concedido e os textos completos do Folha de São Paulo de 1994 foram obtidos. Os textos são armazenados em 365 arquivos em formato SGML (Standard Generalized Markup Language), correspondentes à edição do jornal de cada dia do ano. O enxerto abaixo, referente ao primeiro artigo de 01/01/1994, mostra o formato do conteúdo dos arquivos. Como pode-se ver, cada documento é composto por um identificador, a data do artigo, e o texto do artigo em questão.

```
<DOC>
<DOCNO>FSP940101-001</DOCNO>
<DOCID>FSP940101-001</DOCID>
<DATE>940101</DATE>
<TEXT>
Pesquisa Datafolha feita nas dez principais capitais do país, após um ano de
mandato, aponta o prefeito de Recife, Jarbas Vasconcelos (PMDB), como o mais
popular (63% de ótimo e bom), seguido de Tarso Genro (PT), de Porto Alegre,
com 55% de aprovação. Os prefeitos César Maia (PMDB), do Rio, com 50% de ruim
e péssimo; Lídice da Mata (PSDB), de Salvador, com 48%, e Paulo Maluf (PPR),
de São Paulo, com 41%, tiveram a pior avaliação. A PF vai investigar
pagamentos da Prefeitura de São Paulo a construtoras citadas no caso
Paubrasil. Brasil e Cotidiano
</TEXT>
</DOC>
```

Figura 3.3. Primeiro artigo do arquivo de 01/01/1994

Fonte: Acervo da coleção CHAVE

3.2. Preparação dos *Corpora*

A preparação dos *corpora* de comparação se deu em duas fases elementares: união do conteúdo em um arquivo único e limpeza dos *corpora*. O *corpus* de laudos, no entanto, exigiu fases adicionais para seleção dos laudos.

3.2.1. *Corpora* de Comparação

A primeira fase foi feita com a criação de um *script* com NLTK, que busca o conteúdo desejado nos arquivos, considerando seus formatos, separa por frases e as adiciona a um arquivo único, com uma sentença por linha. Os códigos estão nos Apêndices A e B.

Pela quantidade de artigos, é inviável usar todos no novo *corpus* da Folha. Além disso, a diferença de tamanho entre ele o CorpusDT seria imensa. Decidiu-se então selecionar um artigo por edição do jornal, em um total de 365 artigos. Inicialmente, a escolha do artigo seria feita de maneira aleatória porém, para que a criação do *corpus* possa ser reproduzida, escolheu-se usar o primeiro artigo de cada arquivo.

A segunda fase se deu parcialmente de forma manual, com a checagem de cada *corpus* e a remoção e/ou correção de sentenças e/ou caracteres que afetavam a integridade do *corpus*. Para o *corpus* acadêmico, isso inclui:

- frases separadas por ponto e vírgula (com exceção da listagem simples de itens), como:

  > "Desenvolver sistemas dessa forma tem seus problemas: na ânsia de aproveitar algo já pronto, muitas vezes o sistema resultante não fica tão eficiente; existe a tendência de colar "remendo sobre remendo", produzindo sistemas difíceis de manter; as alterações feitas são tantas que não resta nada do sistema original (nesse caso, talvez tivesse sido melhor partir "do zero"); é também difícil saber qual dos sistemas prontos seria a melhor base para o novo sistema, devido à falta de rigor na documentação."

- quebra incorreta de sentença, devido a presença de ponto em siglas ou abreviações;
- espaços duplos;
- troca de ; e : por ponto quando estavam ao final da frase;
- troca de caixa baixa para caixa alta na primeira letra da sentença.

  Já para o *corpus* jornalístico, a limpeza removeu/corrigiu:

- caracteres não codificados;
- identificadores de seção ou página do artigo: Brasil e Cotidiano; Esporte; São Paulo; Carnaval; Copa 94; PÁG.Esp.3;
- nome do colunista;
- linhas sem pontuação final;
- frases em idiomas estrangeiros;

- quebra incorreta de sentença;
- variações de aspas.

3.2.2. *Corpus* de Laudos

Não foram encontrados *corpora* de laudos médicos em português disponíveis, então foi necessário criar o *corpus*.

A partir de uma base de laudos privada, com 90 categorias de exames que englobam cerca de 150 mil laudos médicos, foram selecionados os laudos usados neste estudo.

Um *corpus* de um tipo único de exame poderia criar viés nos resultados, pois laudos costumam seguir um modelo. Assim, decidiu-se criar um *corpus* heterogêneo, com laudos de 20 diferentes tipos de exames médicos.

O primeiro passo para a preparação do *corpus* foi a seleção das 20 categorias com mais amostras, listadas na tabela a seguir.

Tabela 3.2. Categorias de laudos usados

Tipo	**Número de laudos**
Ultrassonografia Transvaginal	13.730
Ultrassonografia de Articulação	10.898
Ultrassonografia de Abdômen Total	9.083
Ultrassonografia de Mamas Bilateral	8.304
Ultrassonografia do Aparelho Urinário	7.635
Ultrassonografia com Doppler Colorido de Vasos	6.151
Ultrassonografia de Tireoide	5.587
Ultrassonografia de Próstata via Abdominal	4.283
Mamografia Bilateral	3.679
Ultrassonografia de Abdômen Superior	2.644
Tomografia do Crânio	2.014
Ultrassonografia Obstétrica	1.926
Ultrassonografia Pélvica Ginecológica	1.801
Radiografia de Joelho: AP e Lateral	1.650
Radiografia de Coluna Lombo-Sacra	1.323
Ultrassonografia de Bolsa Escrotal	1.293
Ultrassonografia Obstétrica Morfológica	1.207
Radiografia de Tórax: PA e Perfil	997
Tomografia de Coluna Lombo-Sacra	888
Radiografia de Pés/Dedos dos Pés	585

Deste conjunto de 85.678 laudos, decidiu-se selecionar 500 para compôr o *corpus* de

laudos para a pesquisa, de forma que seu tamanho fosse similar aos *corpora* de comparação. Ao final, não selecionou-se 500 laudos, mas sim 489, pelas perdas com arredondamentos nos cálculos ao longo do processo.

A contribuição de laudos de cada categoria se deu em proporção ao seu tamanho, como exibido na Tabela 3.3

Tabela 3.3. Contribuição de cada categoria para o *corpus*

Número de Laudos da Categoria	**Representatividade no Conjunto (%)**	**Proporção para 500 Laudos**	**Número de Laudos no Corpus**
13.730	0,160251172996569	80,1255864982843	80
10.898	0,127197180139592	63,5985900697962	64
9.083	0,106013212259857	53,0066061299283	53
8.304	0,096921029902659	48,4605149513294	48
7.635	0,089112724386657	44,5563621933285	45
6.151	0,071792058638157	35,8960293190784	36
5.587	0,065209271925115	32,6046359625575	33
4.283	0,049989495553118	24,9947477765587	25
3.679	0,042939844534186	21,4699222670931	21
2.644	0,030859730619296	15,4298653096478	15
2.014	0,023506617801536	11,753308900768	12
1.801	0,021020565372675	10,5102826863372	11
1.926	0,022479516328579	11,2397581642895	11
1.650	0,019258152617942	9,6290763089708	10
1.293	0,015091388687878	7,54569434393893	8
1.323	0,015441536917295	7,7207684586475	8
1.207	0,014087630430216	7,04381521510773	7
997	0,011636592824296	5,81829641214781	6
888	0,010364387590747	5,18219379537338	5
585	0,006827890473634	2,73115618945354	3
85.678	100%	500	501

Como o valor da contribuição de cada categoria para o *corpus* foi arredondado, o número total de laudos ficou em 501. Foi decidido usar os valores arredondados tais como exibidos, para respeitar os cálculos.

Em vez de escolher esse conjunto de laudos de forma totalmente aleatória, tentou-se buscar uma amostra que fosse o mais representativa possível dos laudos. Para isso, foi feita a clusterização dos laudos de cada categoria, agrupando os objetos similares entre si, e foram selecionados elementos dos diferentes grupos para compôr o *corpus*.

A clusterização foi feita com o uso de um *software* para mineração de dados, o *RapidMiner*[1]. A técnica de agrupamento usada foi o K-Means, que cria K grupos (*clusters*) e, heurística e iterativamente, calcula a distância entre os elementos, mudando os agrupamentos segundo os valores obtidos, até que não haja uma variação significativa. Como o K-Means funciona com valores numéricos e binários e deseja-se clusterizar objetos polinomiais, usou-se um *kernel* baseado no modelo ANOVA (*ANalysis Of VAriance* / Análise da Variância) para estimar as distâncias entre os elementos e os grupos.

O modelo do processo e os valores usados para o ANOVA são exibidos na Figura 3.4.

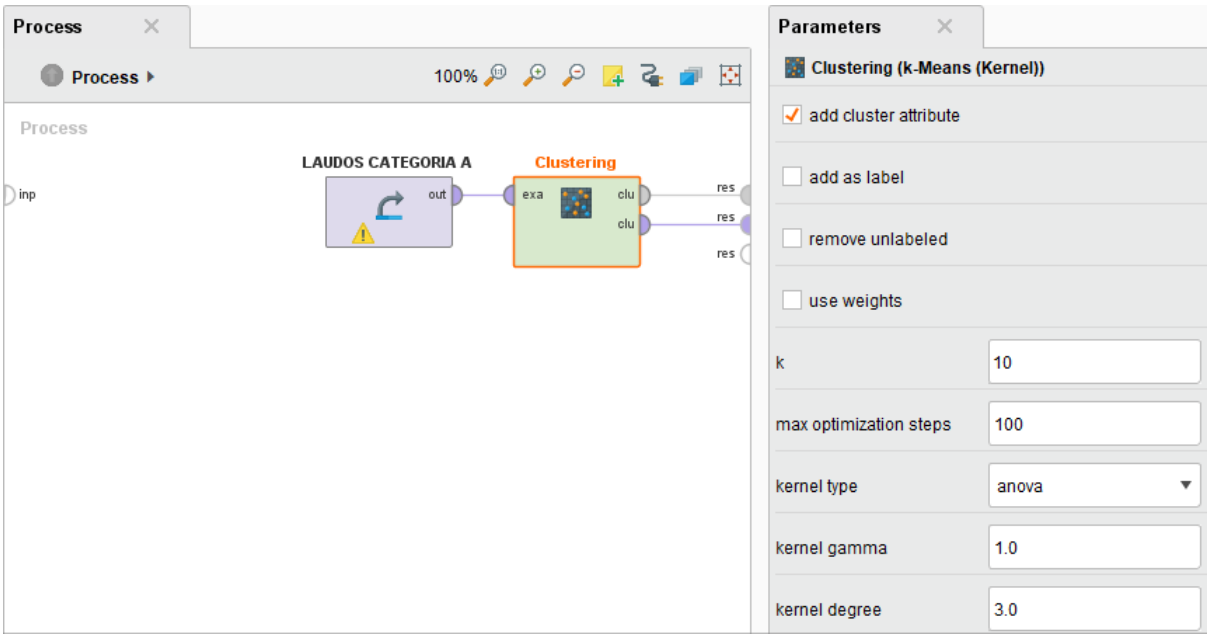

Figura 3.4. Processo e parâmetros da clusterização
Fonte: *Print* da tela da ferramenta.

O número de *clusters* variou de 3 a 10, de acordo com o número de amostras usadas em cada categoria para formar o *corpus* de laudos, como apresentado na Tabela 3.4. O número de laudos a serem retirados de cada *cluster* para compôr o *corpus* varia de acordo com a quantia de elementos em cada grupo, proporcional à representatividade do *cluster* na categoria. Após definido esse número, os laudos foram então selecionados de maneira aleatória.

[1] Acesso em: https://rapidminer.com/

Tabela 3.4. Número de *clusters* para cada categoria de laudos

Número de Laudos da Categoria	Número de Laudos no Corpus	Número de Clusters
585	3	3
888	5	5
997	6	6
1.207	7	7
1.293	8	8
1.323	8	8
1.650	10	10
1.801	11	10
1.926	11	10
2.014	12	10
2.644	15	10
3.679	21	10
4.283	25	10
5.587	33	10
6.151	36	10
7.635	45	10
8.304	48	10
9.083	53	10
10.898	64	10
13.730	80	10

Assim, para a categoria de Tomografia de Coluna Lombo-Sacra, por exemplo, com 888 laudos, que contribui para o *corpus* com 5 laudos, os *clusters* e o número de laudos a serem retirados de cada um são apresentados na Tabela 3.5

Tabela 3.5. Números de laudos de tomografia de coluna lombo-sacra a serem extraídos de cada *cluster*

	Elementos	Representatividade no Grupo (%)	Proporção para 5 Laudos	Laudos Selecionados
Cluster 1	162	18,24324324	0,9121621621	1
Cluster 2	186	20,94594594	1,0472972972	1
Cluster 3	182	20,49549549	1,0247747747	1
Cluster 4	172	19,36936936	0,9684684684	1
Cluster 5	186	20,94594594	1,0472972972	1
Total	888	100%	5	5

Nesse caso, cada *cluster* contribuiu com 1 laudo para o *corpus*, pois seus tamanhos foram homogêneos, devido, provavelmente, à conformidade dos laudos da categoria. O mesmo não se repetiu com todas as categorias, como pode ser visto na Tabela 3.6.

Tabela 3.6. Números de laudos de ultrassonografia de aparelho urinário a serem extraídos de cada *cluster*

	Elementos	**Representatividade no Grupo (%)**	**Proporção para 45 Laudos**	**Laudos Selecionados**
Cluster 1	205	2,68500327	1,2082514734	1
Cluster 2	56	0,73346430	0,3300589390	0
Cluster 3	4.349	56,96136214	25,632612966	26
Cluster 4	53	0,69417157	0,3123772102	0
Cluster 5	47	0,61558611	0,2770137524	0
Cluster 6	676	8,85396201	3,9842829076	4
Cluster 7	60	0,78585461	0,3536345776	0
Cluster 8	678	8,88015717	3,9960707269	4
Cluster 9	697	9,12901113	4,1080550098	4
Cluster 10	814	10,66142763	4,7976424361	5
Total	7;635	100%	45	44

Dada a discrepância no tamanho dos *clusters*, alguns sequer contribuíram com laudos para a formação do *corpus*. Além disso, devido ao arredondamento, o número de laudos a serem selecionados não equivaleu ao valor calculado a *priori*, de 45 laudos. Nos casos em que isso ocorreu, optou-se por usar a quantia determinada pelos arredondamentos, para não beneficiar nem prejudicar qualquer *cluster*.

As tabelas das demais categorias, com o tamanho dos *clusters* e a contribuição de cada um para o *corpus*, encontram-se no Apêndice C.

3.2.3. Estado Final

Ao final, os *corpora* a serem avaliados ficaram com as seguintes características:

Tabela 3.7. Dados dos *corpora* usados no estudo

Corpus	**Tamanho (bytes)**	**Documentos**	**Frases**	**Palavras**
DT	361.992	52	1.945	52.131
Laudos	382.825	489	6.478	50.500
Folha	591.825	365	4.613	92.633

3.3. Processamento dos *Corpora*

Tendo os *corpora* prontos, começou-se a análise.

As características morfológicas analisadas são:

- Etiquetas (classes gramaticais);
- Lemas (unidade elementar das quais as palavras derivaram);
- Unigramas (palavras);
- Bigramas (sequência de duas palavras);
- Trigramas (sequência de três palavras).

Para o processamento, foi utilizada uma ferramenta morfológica desenvolvida em um projeto de extensão dentro do grupo de pesquisa no qual este trabalho está inserido. A seção seguinte apresenta maiores detalhes a respeito da ferramenta.

3.3.1. Ferramenta

A ferramenta usada é um portal web (HTML5/Javascript/CSS3) chamado MorfoX[2], que tem o Freeling como etiquetador. Em testes feitos por Gamallo e Garcia (2015), o Freeling obteve precisão de 96.62% no *corpus* Bosque, 96.99% em um *corpus* de notícias brasileiras, e 96.13% em um *corpus* de artigos da Wikipédia.

O TreeTagger também foi considerado e teve desempenho muito similar em testes mas, pelo Freeling ter código aberto, ele foi escolhido.

O código em C++ criado com as bibliotecas do Freeling faz a etiquetagem e lematização do texto, e provê a lista com as etiquetas, lemas e palavras presentes, tal como sua frequência no *corpus*. Esse código foi traduzido para Javascript com o Emscripten[3], um tradutor de LLVM para Javascript.

O levantamento das demais características (bigramas e trigramas) é feito diretamente no Javascript.

Assim, dado um *corpus* informado pelo usuário, com uma sentença por linha, o portal exibe o número de linhas, de palavras, o tamanho em bytes do arquivo, além do número de etiquetas, lemas, unigramas, bigramas e trigramas únicos, e a lista completa de tais elementos, juntamente com o número de ocorrências no *corpus*.

3.3.1.1. Etiquetas

O *tagset* do Freeling inclui as categorias: adjetivo (A), conjunção (C), determinante (D), substantivo (N), pronome (P), advérbio (R), adposição (S), verbo (V), numeral (Z), data (W), interjeição (I) e pontuação (F).

[2] Projeto MorfoX: https://github.com/Gabrielle7/MorfoX
[3] Emscripten: https://kripken.github.io/emscripten-site

As etiquetas se baseiam nas propostas do grupo EAGLES (*Expert Advisory Group on Language Engineering Standards*) para anotação morfossintática das línguas europeias (LEECH; WILSON, 1996). Nesse modelo, elas têm tamanho variado, sendo que o primeiro caractere especifica a classe gramatical. Um adjetivo, por exemplo, começa com "A" e tem até cinco caracteres, como definido na Tabela 3.8.

Tabela 3.8. Etiquetas para adjetivos

Posição	Atributo	Valor
0	Classe	**A**: Adjetivo
1	Tipo	**O**: ordinal; **Q**: qualificativo; **P**: possessivo
2	Grau	**S**: superlativo; **A**: aumentativo; **C**: diminutivo
3	Gênero	**F**: feminino; **M**: masculino; **C**: comum
4	Número	**S**: singular; **P**: plural; **N**: invariável

Fonte: Adaptado da documentação do FreeLing (TALP-UPC, 2016)

Assim, a palavra "inteligentíssimo" receberia a etiqueta AQSMS (adjetivo qualificativo superlativo masculino singular).

Se um atributo não se aplica ou é irrelevante para a palavra, usa-se "0" (zero) no respectivo campo. Por exemplo, "forte" está no grau normal, então sua etiqueta seria AQ0CS.

O conjunto completo de etiquetas e seus atributos estão descritos em TALP-UPC (2016).

Neste estudo, a etiqueta de pontuação foi desconsiderada, por não agregar informação morfológica ao texto.

3.4. Síntese

Sintetizando a metodologia usada, o primeiro passo foi o levantamento teórico da área, assim como dos *corpora* e etiquetadores disponíveis para português do Brasil, apresentados no Capítulo 2. A seguir, foram selecionados os *corpora* a serem usados no estudo, buscando textos bem formatados. O CorpusDT e o CETENFolha foram escolhidos mas, devido a problemas no segundo, ele foi substituído por um *corpus* similar, formado por uma nova extração de artigos do Folha de São Paulo de 1994. Ambos foram pré-processados (segmentação em frases e limpeza de elementos indesejados) e então criou-se o *corpus* de laudos, com 20 categorias de exames, a partir de uma base de dados privada. Os três *corpora* foram então processados em uma ferramenta WEB (que usa o Freeling), criada em um projeto paralelo. A ferramenta tem como métrica a frequência das características no texto, e extrai as seguintes informações: número de linhas e de palavras, tamanho do arquivo, número de etiquetas, lemas, unigramas, bigramas e trigramas únicos, e a lista completa desses elementos.

Capítulo 4

Resultados e Discussão

O processamento do *corpus* de laudos e dos *corpora* de comparação permitiu transformar tais textos em números e estatísticas que os representam. A partir da lista de etiquetas morfológicas presentes no texto, assim como sua frequência, foi possível ver a composição gramatical de cada *corpus*, enquanto que as listas dos lemas e conjuntos de palavras (unigramas, bigramas e trigramas) permitiram traçar um perfil da composição das frases.

Não há como comparar os resultados obtidos com de outros trabalhos, pois não foram encontrados na literatura estudos que fizessem uma análise e comparação morfológica similares às realizadas neste trabalho.

O resultado do processamento e as estatísticas levantadas para cada característica são apresentados nas seções seguintes.

4.1. Elementos Distintos

Além dos elementos em si (etiquetas, lemas e N-gramas) e quantas vezes aparecem no texto, o MorfoX apresenta ainda o número de elementos distintos encontrados para cada característica. Tais valores são apresentados na Tabela 4.1.

Tabela 4.1. Número de elementos distintos

Corpus	**Etiquetas**	**Lemas**	**Unigramas**	**Bigramas**	**Trigramas**
DT	132	3.579	5.942	26.167	39.636
Folha	156	7.810	12.972	53.082	74.878
Laudos	69	1.956	2.613	7.269	9.693

Esses valores já indicam que os laudos apresentam uma variedade de palavras menor que os outros dois tipos de textos pois, apesar do *corpus* de laudos e do CorpusDT serem

compostos por um número similar de palavras, os laudos tiveram metade do número de etiquetas morfológicas, menos da metade do número de palavras únicas, 3 vezes menos bigramas distintos e 4 vezes menos trigramas.

Por outro lado, o número de frases nos laudos é mais do que o triplo que no *corpus* acadêmico, o que indica que as sentenças são menores nos laudos.

4.2. Classes Gramaticais

Esta seção apresenta cada *corpus* segundo as classes gramaticais encontradas no texto e sua frequência. Atributos como modo, tempo, número e grau não são considerados, apenas as classes em si.

A representatividade de cada classe gramatical no *corpus* acadêmico pode ser vista na Figura 4.1.

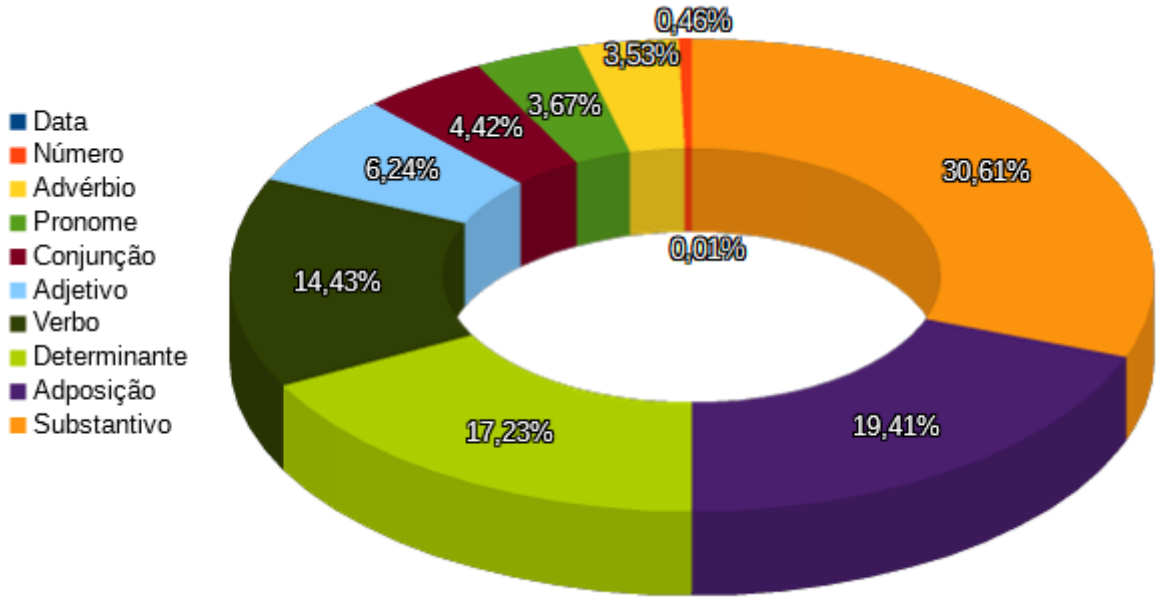

Figura 4.1. Composição do CorpusDT
Fonte: Elaborado pela autora.

Os valores correspondentes à proporção de cada classe gramatical estão explicitados na Tabela 4.2.

Tabela 4.2. Composição gramatical do CorpusDT

Classe	Frequência	Representatividade no Total (%)
Interjeição	0	0
Data	3	0,005572479381826
Número	246	0,456943309309755
Advérbio	1.903	3,53480942120514
Pronome	1.976	3,67040641949625
Conjunção	2.378	4,41711865666097
Adjetivo	3.358	6,23746192139089
Verbo	7.769	14,4308641058028
Determinante	9.277	17,2319637417342
Adposição	10.448	19,4070881937737
Substantivo	16.478	30,6077717512445

Para o *corpus* jornalístico, a representatividade das classes gramaticais é exibida na Figura 4.2.

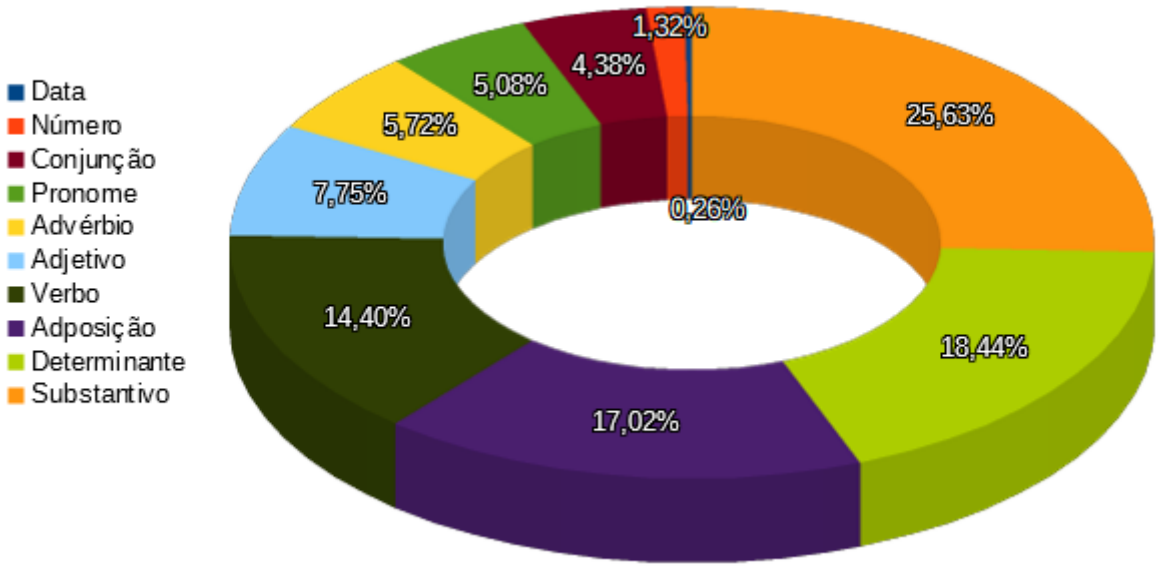

Figura 4.2. Composição do *corpus* Folha
Fonte: Elaborado pela autora.

Os valores estão explicitados na Tabela 4.3.

Tabela 4.3. Composição gramatical do *corpus* Folha

Classe	Frequência	Representatividade no Total (%)
Interjeição	4	0,004102185439293
Data	250	0,256386589955799
Número	1.290	1,32295480417192
Advérbio	4.270	4,37908295644505
Pronome	4.956	5,08260775928376
Conjunção	5.580	5,72254868781343
Adjetivo	7.554	7,74697720210442
Verbo	14.037	14,3955942528382
Determinante	16.594	17,0179162949061
Adposição	17.981	18,4403490959809
Substantivo	24.993	25,6314801710611

Já para o *corpus* de laudos, a representatividade das classes gramaticais nos textos está na Figura 4.3

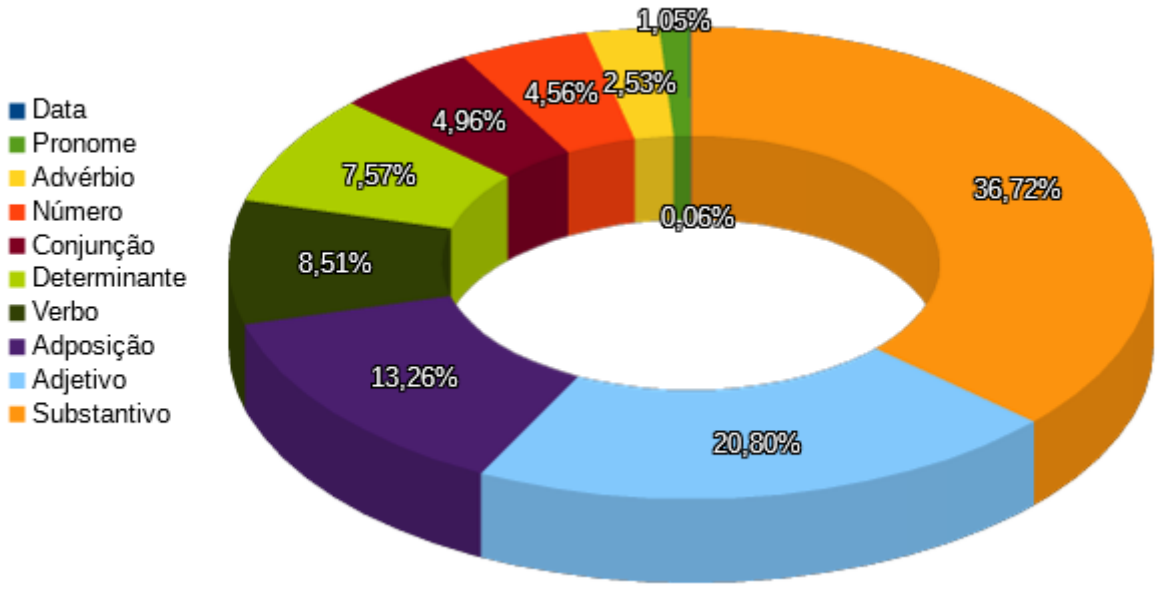

Figura 4.3. Composição do *corpus* de laudos
Fonte: Elaborado pela autora.

Os valores estão explicitados na Tabela 4.4.

Tabela 4.4. Composição gramatical do *corpus* de laudos

Classe	Frequência	Representatividade no Total (%)
Interjeição	0	0
Data	31	0,059193064864142
Número	548	1,04638063050161
Advérbio	1.324	2,52811670581047
Pronome	2.389	4,56168490194955
Conjunção	2.596	4,9569418189456
Adjetivo	3.962	7,56525558037845
Verbo	4.456	8,50852571079414
Determinante	6.942	13,2554276221573
Adposição	10.891	20,7958603043669
Substantivo	19.232	36,7226136602318

Comparando a composição gramatical dos três *corpora*, tem-se a situação apresentada na Figura 4.4

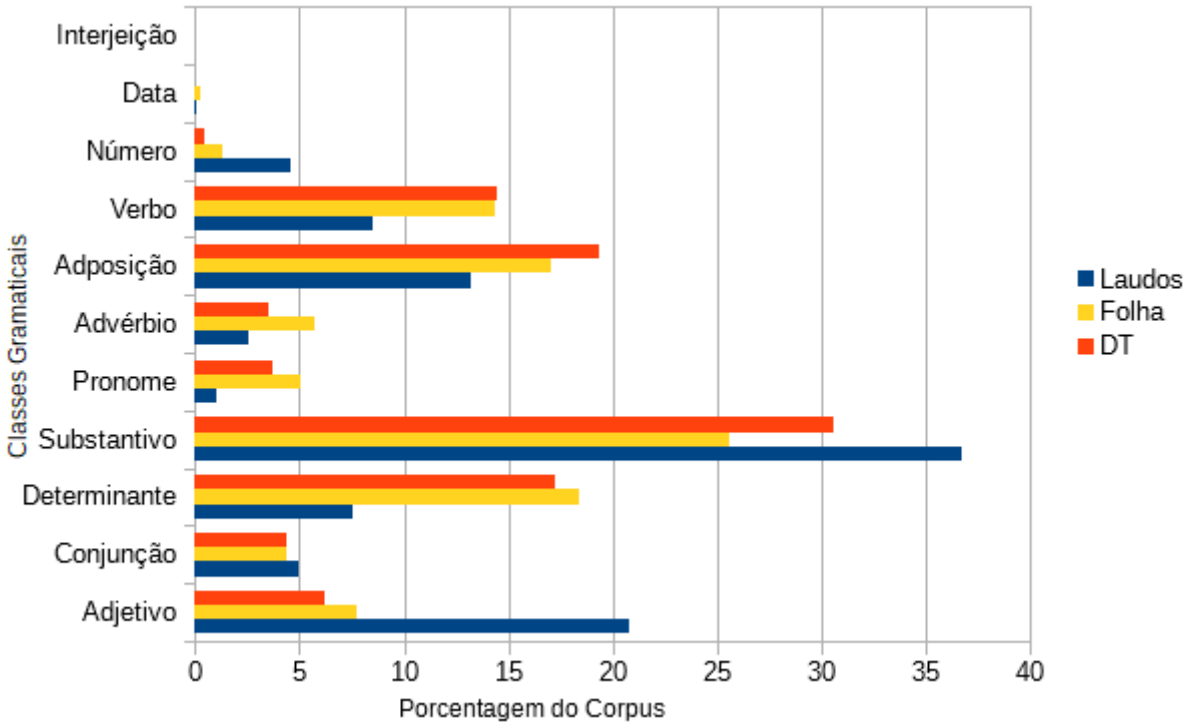

Figura 4.4. Composição dos *corpora*: comparação

Fonte: Elaborado pela autora.

É possível ver que a proporção de números e adjetivos foi muito maior no *corpus* de laudos que nos *corpora* acadêmico e jornalístico.

Para melhor caracterizar a grandeza dessas variações, a Figura 4.5 mostra a diferença da proporção de cada classe gramatical nos *corpora* de comparação em relação aos laudos, em porcentagem. Um valor positivo indica que a presença da classe foi maior no *corpus* de laudos, e valores negativos indicam que foi maior no *corpus* de comparação.

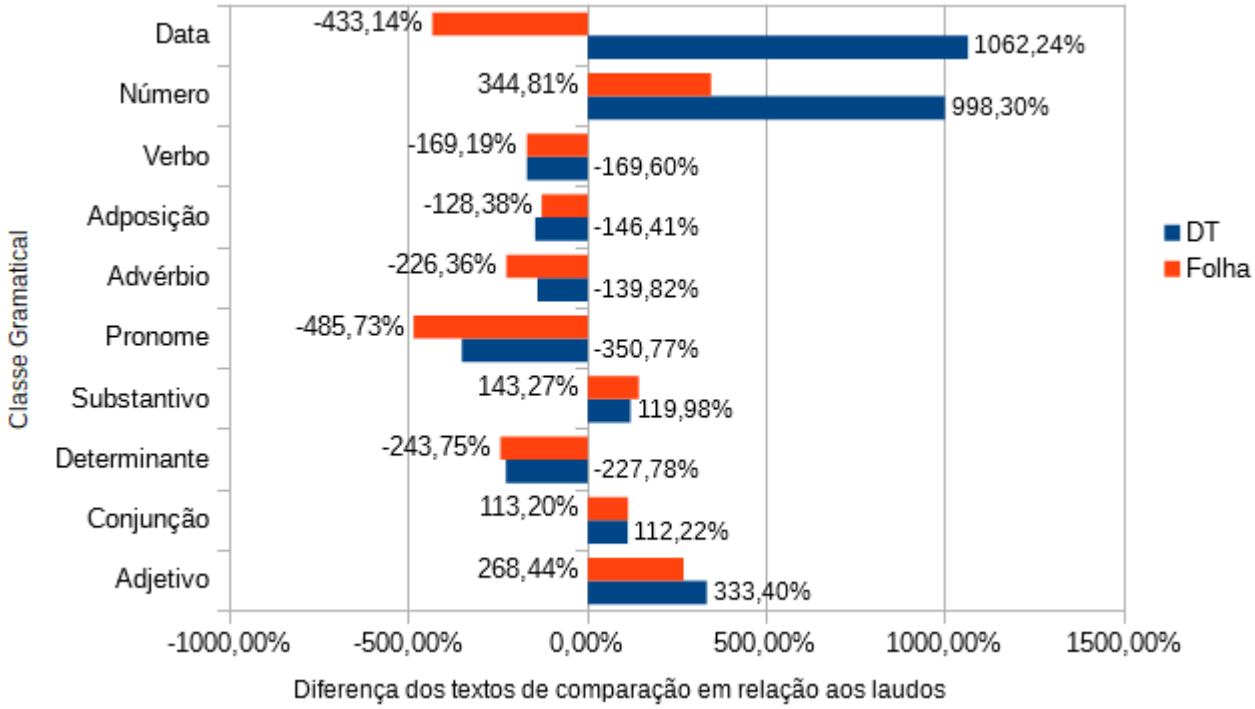

Figura 4.5. Composição dos *corpora*: diferença porcentual

Pode-se ver que os laudos tiveram quase 10 vezes mais números que o *corpus* acadêmico e 3 vezes mais que o *corpus* jornalístico. O *corpus* de laudos também teve cerca de 3 vezes mais adjetivos que os demais. A proporção de substantivos também foi superior, mas em uma escala menor: no *corpus* acadêmico sua presença foi 120% menor que no *corpus* de laudos, e no *corpus* do Folha foi 143% menor.

Conjunção foi a classe mais equilibrada. Teve maior presença nos laudos, mas com variação de apenas 0,5% entre a proporção absoluta dos três *corpora*.

As classes gramaticais com representatividade menor no *corpus* de laudos que nos demais foram: verbos, adposições, advérbios, pronomes e determinantes. Tais dados comprovam a segunda parte das observações, pois isso mostra que as sentenças nos laudos possuem menos conectores, que servem para estabelecer um elo entre as orações que compõem as frases ou entre as próprias frases (PACHECO, 2016), o que sugere sentenças simples e independentes.

A presença de verbos nos laudos não chegou nem a 9% do seu conteúdo, enquanto que nos demais *corpora* chegou a 14,4%, uma diferença de 69%. A presença de determinantes foi ainda menor, com menos de 8%, oposto do que aconteceu nos *corpora* acadêmico e

jornalístico, que tiveram mais determinantes do que verbos. Ambos apresentaram 2 vezes mais determinantes que os laudos.

Adposição foi a quarta classe gramatical com maior presença nos laudos, mas sua proporção ainda foi menor que nos outros dois *corpora*, tendo 68% do número de adposições presentes no CorpusDT e 78% da quantia presente no *corpus* Folha.

Pronomes e advérbios tiveram presenças semelhantes nos *corpus* de comparação, mas o *corpus* de laudos teve mais do que o dobro de advérbios que de pronomes. Em relação ao *corpus* acadêmico, os laudos tiveram 3,5 vezes menos pronomes, e quase 5 vezes menos que no *corpus* jornalistico. Para os advérbios essa diferença foi menor: 1,4 vezes menos que no DT e 2,26 vezes menos que no Folha.

Quanto às interjeições, sua presença nos *corpora* acadêmico e de laudos foi nula, e quase nula (0,004%) no *corpus* jornalístico, então não influenciam a análise.

Por fim, apesar do *corpus* jornalístico ter tido mais do que 4 vezes o número de datas que nos laudos e 46 vezes mais que no *corpus* acadêmico, sua presença foi de apenas 0,26%. Assim, a característica também se mostra irrelevante para o perfilamento morfológico dos textos.

4.3. Lemas

Devido à quantidade de lemas presentes nos textos, são mostrados apenas os 15 mais frequentes em cada *corpus*, para melhor visualização. Eles estão listados na Tabela 4.5.

Tabela 4.5. Lemas mais frequentes

Corpus DT		*Corpus* Folha		*Corpus* de Laudos	
Lema	**Frequência**	**Lema**	**Frequência**	**Lema**	**Frequência**
o	6.700	o	15.303	o	3.442
de	5.689	de	7.939	de	3.438
em	1.604	em	2.981	e	2.175
e	1.457	que	2.622	normal	1.460
ser	1.298	e	2.031	com	1.221
um	1.253	ser	1.905	em	1.104
que	971	a	1.740	x	692
para	842	um	1.594	não	661
a	638	se	1.261	ecotextura	575
sistema	484	para	1.028	ser	521
se	474	por	879	regular	494
esse	439	não	785	se	494
com	431	com	661	contorno	463
por	405	mais	600	cm	437
este	357	seu	549	seu	423

Quanto aos lemas, ficou visível que, para os 3 tipos de textos, os lemas mais frequentes foram artigos (o, a, um), preposições (de, em, para, com, por) e conjunções (e, se, que). O único verbo presente, para todos, foi "ser", sem surpresa, pois é um dos verbos mais comuns da língua portuguesa. Na lista de unigramas mais frequentes, na seção seguinte, revela-se a conjugação mais usada desse verbo: "é" (presente do indicativo).

Classificando esses lemas de acordo com sua classe gramatical, tem-se:

Tabela 4.6. Lemas classificados gramaticalmente

	DT	**Folha**	**Laudos**
Adposição	5	5	3
Conjunção	3	3	2
Artigo	3	3	1
Pronome	2	1	1
Substantivo	1	0	4
Verbo	1	1	1
Adjetivo	0	0	2
Advérbio	0	2	1

Os lemas também indicam uma maior presença de adjetivos e substantivos nos laudos

do que nos demais *corpora*. Enquanto que o DT teve apenas um substantivo e nenhum adjetivo dentre os 15 lemas mais frequentes, o *corpus* de laudos teve 4 substantivos e 2 adjetivos. O *corpus* do Folha não teve sequer um adjetivo ou substantivo na lista.

A Tabela 4.7 exibe uma nova lista de lemas, desconsiderando adposições, pronomes, artigos e conjunções.

Tabela 4.7. Lemas mais frequentes (sem adposições, pronomes, artigos e conjunções)

Corpus DT		*Corpus* Folha		*Corpus* de Laudos	
Lema	**Frequência**	**Lema**	**Frequência**	**Lema**	**Frequência**
ser	1.298	ser	1.905	normal	1.460
sistema	484	não	785	x	692
teste	333	mais	600	não	661
poder	299	ter	495	ecotextura	575
aplicação	258	poder	394	ser	521
software	254	governo	331	regular	494
utilizar	246	haver	328	contorno	463
ter	244	estar	297	cm	437
desenvolvimento	208	país	261	preservar	416
informação	204	fazer	252	espessura	415
ferramenta	193	já	232	forma	389
trabalho	191	ainda	225	dimensão	387
ir	191	mesmo	211	apresentar	380
técnica	189	novo	191	medir	354
não	186	político	189	volume	345

A Tabela 4.8 classifica essa nova lista gramaticalmente.

Tabela 4.8. Lemas classificados gramaticalmente (sem adposições, pronomes, artigos e conjunções)

	DT	**Folha**	**Laudos**
Substantivo	9	4	8
Verbo	5	6	4
Adjetivo	0	1	2
Advérbio	1	4	1

Para o *corpus* Folha, “mais” foi classificado como advérbio, mas pode assumir também função de conjunção ou adjetivo; “mesmo” foi classificado como substantivo, mas pode ser também adjetivo ou conjunção.

Nessa lista refinada de lemas, sem adposições, pronomes, artigos e conjunções, o verbo “ser” assumiu a primeira posição tanto para o DT quanto para o Folha, mas ficou em 5º lugar nos laudos. Como mostrado na Tabela 4.8, as classes gramaticais dos lemas

do *corpus* de laudos e do DT foram similares, com exceção do número de adjetivos, que os laudos tiveram 2, e o DT não teve nenhum. Quem mais destoou foi o *corpus* do Folha, com 4 advérbios e 4 substantivos.

Pelos lemas, consegue-se identificar, ao menos parcialmente, a área de cada texto. Para o CorpusDT, "sistema", "software", "desenvolvimento", "informação" e "ferramenta" já indicam a área de computação. Para o *corpus* Folha, essa visão é mais limitada, pois a presença dos termos "governo", "país" e "político" sugere um texto de cunho político apenas. Para o *corpus* de laudos, "ecotextura", "regular", "contorno", "espessura", "dimensão" e "volume", indicam um texto técnico, com muitas medidas.

4.4. N-Gramas

Para unigramas, bigramas e trigramas também são exibidos os 15 mais frequentes em cada *corpus*. Pontuações foram desconsideradas, mas elementos conectores e acentos foram mantidos, pois são relevantes na análise morfológica dos conjuntos.

4.4.1. Unigramas

As 15 palavras mais frequentes em cada *corpus* foram:

Tabela 4.9. Unigramas mais frequentes

Corpus DT		*Corpus* Folha		*Corpus* de Laudos	
Palavra	**Frequência**	**Palavra**	**Frequência**	**Palavra**	**Frequência**
de	5.638	de	7.815	de	3.414
a	2.920	a	6.477	e	2.175
o	2.465	o	6.056	a	1.218
e	1.441	em	2.713	com	1.173
em	1.431	que	2.613	em	1.055
que	971	e	1.924	o	995
os	798	os	1.644	normal	911
para	789	as	1.461	x	686
um	672	se	1.200	normais	549
as	651	para	975	os	515
uma	523	um	845	Não	506
é	498	por	833	se	494
se	469	é	797	ecotextura	490
com	390	uma	721	contornos	447
por	386	não	691	cm	4373

A classificação gramatical dessas palavras é mostrada na Tabela 4.10.

Tabela 4.10. Unigramas classificados gramaticalmente

	DT	**Folha**	**Laudos**
Adposição	5	4	3
Conjunção	3	3	2
Artigo	6	6	3
Pronome	0	0	0
Substantivo	0	0	4
Verbo	1	1	0
Adjetivo	0	0	2
Advérbio	0	1	1

Os unigramas gerais, como era esperado, apresentaram as mesmas características dos lemas, o domínio de adposições, conjunções e artigos, em uma proporção ainda maior para os *corpora* de comparação.

O *corpus* de laudos foi o único a apresentar adjetivos (dois) e substantivos (quatro) na lista de palavras mais frequentes, e nenhum verbo, o que, novamente, reforça as observações empíricas. Além disso, teve metade do número de artigos que os demais, sendo que nenhum era artigo indefinido (um, uma, uns, umas). Essa é uma questão relevante ao formato dos laudos, pois o uso de artigos definidos (o, a, os as) indica com precisão o substantivo, enquanto que artigos indefinidos referenciam simplesmente um elemento de uma categoria (PEREZ, 2016). Isso sugere que nos laudos os elementos são citados com precisão e especificidade.

Assim como feito com os lemas, a tabela seguinte desconsidera adposições, pronomes, artigos e conjunções da lista dos unigramas mais frequentes.

Tabela 4.11. Unigramas mais frequentes (sem adposições, pronomes, artigos e conjunções)

Corpus DT		*Corpus* Folha		*Corpus* de Laudos	
Palavra	**Frequência**	**Palavra**	**Frequência**	**Palavra**	**Frequência**
é	498	é	797	normal	911
teste	305	não	691	x	686
ser	266	mais	570	normais	549
sistemas	254	governo	308	Não	506
software	247	ser	264	ecotextura	490
sistema	225	É	225	contornos	447
são	216	país	225	cm	437
desenvolvimento	205	ainda	216	regulares	425
mais	190	foi	203	espessura	391
não	184	já	200	é	329
trabalho	154	inflação	189	mm	318
aplicações	137	Brasil	183	dimensões	311
informações	137	presidente	159	alterações	310
hipermídia	133	há	158	parênquima	299
técnicas	125	pode	149	Ausência	297

A classificação gramatical das palavras dessa nova lista é mostrada na Tabela 4.12.

Tabela 4.12. Unigramas classificados gramaticalmente (sem adposições, pronomes, artigos e conjunções)

	DT	**Folha**	**Laudos**
Substantivo	10	5	10
Verbo	3	7	1
Adjetivo	1	0	3
Advérbio	1	3	1

Novamente, “mais” foi classificado como advérbio, mas pode assumir também função de conjunção ou adjetivo.

Mais uma vez os laudos apresentaram mais adjetivos e menos verbos, em valores consideráveis, enquanto que o *corpus* jornalístico foi o que teve mais unigramas verbais e menos substantivos.

4.4.2. Bigramas

A Tabela 4.13 mostra os 15 bigramas com mais ocorrências em cada *corpus*.

Tabela 4.13. Bigramas mais frequentes

Corpus DT		*Corpus* Folha	
Bigrama	**Frequência**	**Bigrama**	**Frequência**
de teste	237	que o	375
de software	190	que a	234
de um	172	que se	193
para a	139	de que	174
para o	110	para o	172
desenvolvimento de	109	o que	161
de dados	109	de um	148
de uma	100	para a	143
com o	94	e a	139
teste de	88	o governo	131
de sistemas	87	com a	131
e a	87	de uma	117
do sistema	78	com o	117
análise de	78	e o	113
processo de	77	do que	109

Corpus de Laudos	
Bigrama	**Frequência**
contornos regulares	296
ausência de	296
sem alterações	241
espessura normal	241
não há	206
de calibre	202
dimensões normais	201
calibre normal	189
impressão diagnóstica	185
a ecotextura	184
não se	173
sinais de	170
e dimensões	163
do parênquima	159
há sinais	150

Nos bigramas, para os *corpora* de comparação, quase todas as instâncias possuem

elementos de conexão, ou seja, preposições e conjunções, assim, não possuem um sentido, pois faltam os elementos sendo conectados. Apenas uma instância (única do *corpus* do Folha que não é composta por artigos, adposições e conjunções), "o governo", possui uma combinação significativa. No *corpus* de laudos, por outro lado, apenas 5 elementos possuem conectores e, no total, 8 possuem significância, 5 delas caracterizando o estado de um elemento ("contornos regulares", "sem alterações", "espessura normal", "dimensões normais", "calibre normal").

Esse fato também reforça a análise de que os *corpora* acadêmico e jornalístico possuem sentenças mais complexas, pela repetição de conectores, enquanto que o *corpus* de laudos é mais conciso, com menos orações em cada frase.

Além disso, enquanto que os laudos contaram com 12 substantivos e 3 adjetivos na lista dos 15 bigramas mais frequentes, o CorpusDT apresentou 9 substantivos e nenhum adjetivo, e o *corpus* do Folha não teve nem um substantivo.

4.4.3. Trigramas

Por fim, a Tabela 4.14 mostra os 15 trigramas com mais ocorrências.

Todas as instâncias exibidas do CorpusDT apresentaram um conector, a preposição "de". Dessas, 9 apresentam sentido, todas na forma "substantivo de substantivo", como "casos de teste" e "engenharia de software". Pode-se ver que os assuntos presentes no *corpus* agora ficam mais evidentes.

O *corpus* jornalístico ainda apresentou vários trigramas com conectores, mas agora possui também dois substantivos pessoais em destaque: "fernando henrique cardoso" e "luiz inácio lula da silva". Isso se deve diretamente ao período ao qual os textos pertencem, 1994, ano de eleição.

O *corpus* de laudos apresentou agora instâncias um pouco mais completas que nos unigramas e bigramas, com mais conjunções e preposições, e com frequências de repetição muito superior aos trigramas encontrados nos demais *corpora*. Essa repetição indica que os laudos usam termos limitados, e seguem um formato bem definido. É o oposto do *corpus* Folha que, ao apresentar repetição de conectores mas não dos elementos sendo conectados, indica que esses elementos são variados.

Tabela 4.14. Trigramas mais frequentes

Corpus DT	
Trigrama	**Frequência**
análise de mutantes	59
atividade de teste	53
casos de teste	53
o desenvolvimento de	49
de teste de	42
a atividade de	39
o processo de	37
desenvolvimento de software	35
engenharia de software	34
critério análise de	34
um conjunto de	33
de casos de	31
teste de regressão	30
o objetivo de	29
a necessidade de	29

Corpus Folha	
Trigrama	**Frequência**
fernando henrique cardoso	58
de que o	41
de são paulo	33
é evidente que	32
que o governo	29
luiz inácio lula	29
inácio lula da	29
lula da silva	29
o fato de	26
de que a	24
é claro que	24
mais do que	24
em vez de	23
o plano real	23
não se pode	22

Corpus de Laudos	
Trigrama	**Frequência**
de calibre normal	162
e dimensões normais	154
há sinais de	150
não há sinais	147
formas e dimensões	147
não se observam	146
com formas e	120
nos seus maiores	117
seus maiores eixos	117
exame realizado com	109
a ecotextura do	103
de líquido livre	91
líquido livre na	91
mm de espessura	89
imagens sugestivas de	89

4.5. Questão de Pesquisa

Respondendo à questão de pesquisa, de forma resumida, as características morfológicas que se mostraram predominantes em laudos médicos foram a grande quantidade de adjetivos e substantivos, uma presença pequena de elementos conectores, como preposições e conjunções, o que sugere frases mais simples, além de uma limitação na variedade de palavras usadas. Bigramas e trigramas indicaram um texto descritivo, com muitas medidas, e uma variação menor na combinação das palavras, o que sugere que os laudos seguem um padrão mais definido que os demais tipos de textos. A quantidade de números foi muito maior que nos *corpora* de comparação, enquanto que a quantidade de verbos, pronomes e determinantes foi bem menor.

Capítulo 5

Conclusão

Algumas características morfológicas de laudos médicos podem ser vistas empiricamente, como o alto grau de substantivos e adjetivos, e a presença de sentenças gramaticalmente simples.

Contudo, dados não comprovados por métodos científicos não são o suficiente para embasar pesquisas científicas. Assim, a presente monografia buscou prover esse embasamento, analisando um *corpus* de laudos e fazendo o levantamento do perfil morfológico deles. A pesquisa ainda foi além, mostrando também as diferenças do conteúdo de laudos para com outros tipos de textos, considerando os elementos: classes gramaticais, lemas, unigramas, bigramas e trigramas.

A área de morfologia, a etiquetagem de *corpus* e o uso de *corpora* foram pesquisados para embasar este estudo. A seguir, foram selecionados dois *corpora*, para servirem de comparação, um jornalístico e um acadêmico. Eles foram formatados e preparados para serem processados. O *corpus* de laudos médicos foi criado, contendo 489 laudos, de 20 tipos diferentes de exames (escolheu-se as categorias com mais laudos, dentre 90 tipos). Por fim, os três *corpora* foram processados. A ferramenta usada gerou a frequência das etiquetas presentes nos textos, assim como das demais características, e a lista completa dos elementos. Essas informações foram analisadas e levantou-se estatísticas em cima delas, que representassem os *corpora*.

Os resultados comprovaram as hipóteses de pesquisa observadas empiricamente por pesquisadores. Como pôde-se ver, os laudos, de fato, apresentaram uma quantidade maior de substantivos, números e adjetivos que os demais tipos de textos, em proporções consideráveis. Isso foi visto nas análises das 5 características estudadas. As demais classes gramaticais, verbos, adposições, advérbios, pronomes e determinantes, tiveram uma presença consideravelmente menor nos laudos do que nos outros dois *corpora*.

A discussão também levantou diversos indícios da simplicidade e concisão das frases

que compõem os laudos. O número de palavras distintas presentes no *corpus* de laudos, em comparação com o dos demais, indica um vocabulário limitado. Além disso, a baixa proporção de conectores, como conjunções e preposições, indica que as sentenças são reduzidas, com poucas orações. A ausência de artigos indefinidos indica que os elementos nos laudos são referenciados com especificidade, o que reforça a ideia da presença de sentenças diretas e concisas. Os N-gramas encontrados nos laudos mostram ainda que a linguagem é descritiva, razão da presença de tantos adjetivos e números.

Este trabalho traduz observações empíricas em dados estatísticos a respeito da morfologia de textos de laudos médicos. Tais dados possibilitam aos pesquisadores avaliar se suas técnicas de recuperação de informação são adequadas para textos com o perfil levantado. Além disso, podem embasar pesquisas da estrutura de textos de laudos não só quanto à morfologia, mas quanto às demais etapas do Processamento de Linguagem Natural (análises sintática e pragmático-discursiva).

Em pesquisas futuras, *corpora* de outras categorias poderiam ser avaliados, para ver se as distinções permanecem, são amenizadas ou acentuadas. O *corpus* de laudos também poderia variar. Um *corpus* criado de forma totalmente aleatória apresentaria os mesmos resultados? A variação do número de amostras mudaria algo? Essas questões poderiam enriquecer o perfil morfológico de laudos e suas diferenças para com outros tipos de textos, levantados por esta monografia.

Referências

AIRES, R. V. X. et al. Combining Multiple Classifiers to Improve Part of Speech Tagging: A Case Study for Brazilian Portuguese. In: *Brazilian Symposium on Artificial Intelligence (SBIA'2000)*. Atibaia, SP, Brasil: [s.n.], 2000. p. 20–22.

ALENCAR, L. F. Aelius: uma ferramenta para anotação automática de corpora usando o NLTK. In: *IX Encontro de Linguística de Corpus (ELC 2010)*. Porto Alegre, RS, Brasil: [s.n.], 2010.

ALUÍSIO, S. et al. An account of the challenge of tagging a reference corpus for brazilian portuguese. In: *Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language*. Faro, Portugal: [s.n.], 2003.

ALUÍSIO, S. M.; AIRES, R. V. *Etiquetação de um Corpus e Construção de um Etiquetador de Português*. São Carlos, SP, Brazil, 2000. (Relatórios Técnicos do ICMC-USP, NILC-TR-00-2).

ANTIQUEIRA, L.; FELTRIM, V. D.; NUNES, M. G. V. *Projeto e Implementação do Sistema SciPo*. São Carlos, SP, Brazil, 2003. (Relatórios Técnicos do ICMC, 223).

BICK, E. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Tese (Doutorado) — Aarhus University, 2000.

BRILL, E. Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part-of-speech Tagging. *Computational Linguistics*, MIT Press, Cambridge, MA, USA, v. 21, n. 4, p. 543–565, dez. 1995.

FELTRIM, V. D.; NUNES, M. G. V.; ALUÍSIO, S. M. *Um Corpus de Textos Científicos em Português para a Análise da Estrutura Esquemática*. São Carlos, SP, Brazil, 2001. (Série de Relatórios Técnicos do NILC, NILC-TR-01-4).

FRANCIS, W. N.; KUCERA, H. Manual of Information to Accompany a Standard Sample of Present-Day Edited American English for Use with Digital Computers. Brown University, Department of Lingustics, Providence, Rhode Island, USA, 1964. Revisado em 1971. Revisado e ampliado em 1979.

GAMALLO, P.; GARCIA, M. Yet another suite of multilingual NLP tools. In: *Symposium on Languages, Applications and Technologies*. [S.l.: s.n.], 2015. p. 81–90.

HELMUT, S. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK: [s.n.], 1994.

INDURKHYA, N.; DAMERAU, F. J. (Ed.). *Handbook of Natural Language Processing*. 2. ed. [S.l.]: Chapman & Hall/CRC, 2010. (Machine Learning and Pattern Recognition Series).

JURAFSKY, D.; MARTIN, J. H. Language Modeling with N-Grams. In: *Speech and Language Processing (3rd ed. draft)*. [s.n.], 2016. cap. 4. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>.

JURAFSKY, D.; MARTIN, J. H. Language Modeling with N-Grams. In: *Speech and Language Processing (3rd ed. draft)*. [s.n.], 2016. cap. 10. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>.

LEECH, G.; WILSON, A. *Recommendations for the Morphosyntactic Annotation of Corpora*. [S.l.], 1996. (EAGLES TECHREPORT, EAG–TCWG–MAC/R). Disponível em: <www.ilc.cnr.it/EAGLES/annotate/annotate.html>.

LINGUATECA. *CETENFolha*. 2016. Disponível em: <http://www.linguateca.pt/CETENFolha/>.

LUCCA, J. L.; NUNES, M. G. V. *Lematização versus Stemming*. São Carlos, SP, Brazil, 2002. (Relatórios Técnicos do ICMC-USP, NILC-TR-02-22).

MARCUS, M. P.; MARCINKIEWICZ, M. A.; SANTORINI, B. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics - Special issue on using large corpora: II*, Cambidge, MA, USA, v. 19, n. 2, p. 313–330, 1993.

OTHERO, G. A.; AYRES, M. R. Anotação morfológica automática de corpus de língua falada: desafios ao Aelius. *Texto Livre: Linguagem e Tecnologia*, v. 7, n. 2, 2014.

PACHECO, M. do C. *O que são conectivos?* 2016. Disponível em: <http://brasilescola.uol.com.br/o-que-e/portugues/o-que-sao-conectivos.htm>. Acesso em: 27/06/2017.

PADRÓ, L et al. FreeLing 2.1: Five Years of Open-source Language Processing Tools. In: *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*. La Valletta, Malta: [s.n.], 2010.

PARDO, T. A. S. *Manual do ReGra: REvisor GRAmatical*. São Paulo, SP, 2000. (Série de Relatórios do NILC, NILC-TR-00-10).

PEREZ, L. C. A. *O que é artigo?* 2016. Disponível em: <http://brasilescola.uol.com.br/o-que-e/portugues/o-que-e-artigo.htm>. Acesso em: 27/06/2017.

PINHEIRO, G. M.; ALUÍSIO, S. M. *Corpus Nilc: descrição e análise crítica com vistas ao projeto Lacio-Web*. São Carlos, SP, Brazil, 2003. (Série de Relatórios Técnicos do NILC, NILC-TR-03-03).

RATNAPARKHI, A. A Maximum Entropy Part-of-Speech Tagger. *Proceedings of the First Empirical Methods in Natural Language Processing Conference*, Philadelphia, Pa, USA, 1996.

SANTOS, D. O projecto Processamento Computacional do Português: Balanço e perspectivas. In: NUNES, Maria das Graças Volpe (Ed.). *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR)*. São Paulo, Brasil, 2000. p. 105–113.

SANTOS, D. Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. *Linguamática 1.1*, p. 25–59, 2009.

SILVA, B. C. D. et al. *Introdução ao processamento das línguas naturais e algumas aplicações*. [S.l.], 2007. 121 p. (Série de Relatórios do NILC, NILC-TR-07-10).

TALP-UPC. *FreeLing User Manual*. [S.l.], 2016. Disponível em: <https://talp-upc.gitbooks.io/freeling-user-manual/content/>.

VIEIRA, R; LIMA, V. L. S. Linguística Computacionxal: princípios e aplicações. In: NEDEL, Luciana (Ed.). *IX Escola de Informática da SBC-Sul (ERI 2001)*. Porto Alegre, RS, Brasil: [s.n.], 2001. p. 27–61.

WYNNE, M. Developing Linguistic Corpora: a Guide to Good Practice. *Oxford: Oxbow Books*, 2005. Disponível em: <http://ota.ox.ac.uk/documents/creating/dlc/>.

Glossário

- adjetivo: (classe gramatical) Expressa uma característica de um ser. Ex: bonito, escuro, maduro, veloz, divertido, grande.
- adposição: (classe gramatical) Palavras que conectam orações. É formada por preposições (indica subordinação do termo posterior ao termo anterior), posposições (indica subordinação do termo anterior ao termo posterior; não existe no português) e circumposições (circula a frase; é raríssima, existe em poucos idiomas, o que não inclui o português). A morfologia do português não usa essa classe gramatical, mas sim uma de suas subclasses, preposições, já que não há posposições nem circumposições na língua.
- advérbio: (classe gramatical) Palavra que modifica o sentido de um verbo, adjetivo ou outro advérbio. Ex: não, muito, sempre.
- artigo: (classe gramatical) Indica um substantivo de maneira indefinida ou definida. Ex: o, a, os, as, um, uma, uns, umas.

- bigrama: Sequência de duas palavras.

- conjunção: (classe gramatical) Palavras que conectam orações e frases. Ex: embora, porém, se, mas, e, quando.
- corpora: Plural de *corpus*.
- corpus: Conjunto de textos autênticos que compõem uma amostra representativa de uma língua ou variação linguística.

- determinante: (classe gramatical) Determina um nome, antecedendo-o. É formado por artigos, possessivos (meu, seu, nosso, minhas, sua), demonstrativos (este, isso, aquilo, aqueles, essas), indefinidos (certa, qualquer, outro, nenhum, alguma, muitos, todas), interrogativos (qual, quanto, quem, que, quais) e numerais (primeiro, segundo, um, dois). A morfologia do português não usa essa classe gramatical, mas sim as subclasses que a formam.

- interjeição: (classe gramatical) Exprime sensações ou emoções. Ex: Droga!, oh, Bis!, Ai, Psiu, Hum, Ora Bolas, ah.

- laudo: Documento criado por um médico especialista, onde este descreve os elementos observados durante um exame médico.

- lema: Unidade elementar da qual se derivam as palavras. Para verbos usa-se o infinitivo, e para substantivos e adjetivos usa-se o singular masculino.
- morfologia: Estudo da estrutura das palavras e sua classificação gramatical. As palavras são vistas de forma isolada, sem considerar o resto da oração..
- preposição: (classe gramatical) Conecta duas orações. Ex: de, para, por, em, durante, que, sobre.
- pronome: (classe gramatical) Acompanham ou substituem um nome (substantivo, adjetivo, pronome e artigo). Ex: ele, eu, minha, sua, nossa, essa.
- substantivo: (classe gramatical) Dão nome aos seres (objetos, pessoas, fenômenos, ações, estados, lugares). Ex: Brasil, corrida, João, mochila, chuva, guerra, borboleta, alegria.
- trigrama: Sequência de três palavras.
- unigrama: Sequência de uma palavra.
- verbo: (classe gramatical) Determina uma ação ou estado. Ex: falar, ficar, vender, ser, cantar, sou, era, falou, fica, venda, cantaram, corremos.

Apêndices

Apêndice

A

Script para o CorpusDT

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
import os, codecs, re, nltk
from sys import argv

def generate(in_folder, out_filename, tokenizer):
    pular = False
    files = [n for n in os.listdir(in_folder) if n.endswith(".txt")]
    files.sort()
    with open(out_filename, 'w') as fo:
        for file in files:
            with codecs.open(os.path.join(in_folder, file), 'r', 'cp1252') as fi:
                for line in fi:
                    sentences = tokenizer.tokenize(line)
                    for s in sentences:
                        if len(s) > 2:
                            if s[:2] in ("- ", "; "):
                                s = s[2:]
                            if s[-1] in (":", ";"):
                                s = s[:-1]
                            elif s[-3:] == "; e":
                                s = s[:-3]
                        fo.write(s)
                        if s[-1] != '\n':
                            fo.write('\n')

if __name__ == "__main__":
    if len(argv) != 3:
        print("Uso: " + argv[0] + " pasta arquivo_de_saida")
        exit(1)
    try:
        nltk.data.find('tokenizers/punkt/portuguese.pickle')
    except LookupError:
        nltk.download('punkt')
    sent_tokenizer = nltk.data.load('tokenizers/punkt/portuguese.pickle')
    generate(argv[1], argv[2], sent_tokenizer)
```

Apêndice

B

Script para o Corpus do Folha

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
import os, codecs, re, nltk
from sys import argv

def generate(in_folder, out_filename, tokenizer):
    pular = False
    files = [n for n in os.listdir(in_folder) if n.endswith(".sgml")]
    files.sort()
    with open(out_filename, 'w') as fo:
        for file in files:
            with codecs.open(os.path.join(in_folder, file), 'r', 'latin1') as fi:
                inside = False
                for line in fi:
                    if line == "<TEXT>\n":
                        inside = True
                    elif line == "</TEXT>\n":
                        inside = False
                        break
                    elif inside:
                        sentences = tokenizer.tokenize(line)
                        for s in sentences:
                            fo.write(s)
                            if s[-1] != '\n':
                                fo.write('\n')

if __name__ == "__main__":
    if len(argv) != 3:
        print("Uso: " + argv[0] + " pasta arquivo_de_saida")
        exit(1)
    try:
        nltk.data.find('tokenizers/punkt/portuguese.pickle')
    except LookupError:
        nltk.download('punkt')
    sent_tokenizer = nltk.data.load('tokenizers/punkt/portuguese.pickle')
    generate(argv[1], argv[2], sent_tokenizer)
```

Apêndice C

Clusterização dos Laudos

Tabela C.1. Laudos de radiografia de pés/dedos dos pés a serem extraídos de cada *cluster*

	Elementos	Representatividade no Grupo (%)	Proporção para 10 laudos	Laudos Selecionados
Cluster 1	189	32,3076923076923	0,969230769230769	1
Cluster 2	168	28,7179487179487	0,861538461538462	1
Cluster 3	228	38,974358974359	1,16923076923077	1
Total	585	100%	3	3

Tabela C.2. Laudos de radiografia de tórax PA e perfil a serem extraídos de cada *cluster*

	Elementos	Representatividade no Grupo (%)	Proporção para 10 laudos	Laudos Selecionados
Cluster 1	48	4,81444332999	0,288866599799398	0
Cluster 2	302	30,2908726178536	1,81745235707121	2
Cluster 3	145	14,543630892678	0,872617853560682	1
Cluster 4	136	13,6409227683049	0,818455366098295	1
Cluster 5	195	19,5586760280843	1,17352056168506	1
Cluster 6	171	17,1514543630893	1,02908726178536	1
Total	997	100%	6	6

Tabela C.3. Laudos de ultrassom obstétrico morfológico a serem extraídos de cada *cluster*

	Elementos	Representatividade no Grupo (%)	Proporção para 10 laudos	Laudos Selecionados
Cluster 1	174	14,4159072079536	1,00911350455675	1
Cluster 2	169	14,0016570008285	0,980115990057995	1
Cluster 3	189	15,6586578293289	1,09610604805302	1
Cluster 4	172	14,2502071251036	0,997514498757249	1
Cluster 5	168	13,9188069594035	0,974316487158244	1
Cluster 6	163	13,5045567522784	0,945318972659486	1
Cluster 7	172	14,2502071251036	0,997514498757249	1
Total	1207	100%	7	7

Tabela C.4. Laudos de ultrassonografia de bolsa escrotal a serem extraídos de cada *cluster*

	Elementos	Representatividade no Grupo (%)	Proporção para 10 laudos	Laudos Selecionados
Cluster 1	162	12,5290023201856	1,00232018561485	1
Cluster 2	146	11,291569992266	0,903325599381284	1
Cluster 3	146	11,291569992266	0,903325599381284	1
Cluster 4	197	15,2358855375097	1,21887084300077	1
Cluster 5	155	11,9876256767208	0,959010054137664	1
Cluster 6	162	12,5290023201856	1,00232018561485	1
Cluster 7	155	11,9876256767208	0,959010054137664	1
Cluster 8	170	13,1477184841454	1,05181747873163	1
Total	1293	100%	8	8

Tabela C.5. Laudos de radiografia de coluna lombo-sacra a serem extraídos de cada *cluster*

	Elementos	Representatividade no Grupo (%)	Proporção para 10 laudos	Laudos Selecionados
Cluster 1	196	14,8148148148148	1,18518518518519	1
Cluster 2	155	11,7157974300831	0,937263794406652	1
Cluster 3	24	1,8140589569161	0,145124716553288	0
Cluster 4	219	16,5532879818594	1,32426303854875	1
Cluster 5	177	13,3786848072562	1,0702947845805	1
Cluster 6	173	13,0763416477702	1,04610733182162	1
Cluster 7	175	13,2275132275132	1,05820105820106	1
Cluster 8	204	15,4195011337868	1,23356009070295	1
Total	1323	100%	8	7

Tabela C.6. Laudos de radiografia de joelho AP e lateral a serem extraídos de cada *cluster*

	Elementos	Representatividade no Grupo (%)	Proporção para 10 laudos	Laudos Selecionados
Cluster 1	80	4,8484848484849	0,484848484848485	0
Cluster 2	32	1,9393939393939	0,193939393939394	0
Cluster 3	612	37,0909090909091	3,70909090909091	4
Cluster 4	134	8,1212121212121	0,812121212121212	1
Cluster 5	61	3,6969696969697	0,36969696969697	0
Cluster 6	108	6,5454545454546	0,654545454545455	1
Cluster 7	216	13,0909090909091	1,30909090909091	1
Cluster 8	211	12,7878787878788	1,27878787878788	1
Cluster 9	173	10,4848484848485	1,04848484848485	1
Cluster 10	23	1,3939393939394	0,139393939393939	0
Total	1650	100%	10	9

Tabela C.7. Laudos de ultrassonografia pélvica ginecológica a serem extraídos de cada *cluster*

	Elementos	Representatividade no Grupo (%)	Proporção para 10 laudos	Laudos Selecionados
Cluster 1	190	10,5496946141033	1,16046640755136	1
Cluster 2	194	10,7717934480844	1,18489727928928	1
Cluster 3	169	9,3836757357024	1,03220433092726	1
Cluster 4	168	9,3281510272071	1,02609661299278	1
Cluster 5	186	10,3275957801222	1,13603553581344	1
Cluster 6	177	9,8278734036646	1,08106607440311	1
Cluster 7	186	10,3275957801222	1,13603553581344	1
Cluster 8	163	9,0505274847307	0,995558023320378	1
Cluster 9	176	9,7723486951694	1,07495835646863	1
Cluster 10	192	10,6607440310938	1,17268184342032	1
Total	1801	100%	11	10

Tabela C.8. Laudos de ultrassonografia obstétrica a serem extraídos de cada *cluster*

	Elementos	**Representatividade no Grupo (%)**	**Proporção para 10 laudos**	**Laudos Selecionados**
Cluster 1	203	10,539979231568	1,15939771547248	1
Cluster 2	207	10,7476635514019	1,18224299065421	1
Cluster 3	179	9,2938733125649	1,02232606438214	1
Cluster 4	180	9,3457943925234	1,02803738317757	1
Cluster 5	195	10,1246105919003	1,11370716510903	1
Cluster 6	190	9,865005192108	1,08515057113188	1
Cluster 7	201	10,4361370716511	1,14797507788162	1
Cluster 8	180	9,3457943925234	1,02803738317757	1
Cluster 9	191	9,9169262720665	1,09086188992731	1
Cluster 10	200	10,3842159916926	1,14226375908619	1
Total	1926	100%	11	10

Tabela C.9. Laudos de tomografia de crânio a serem extraídos de cada *cluster*

	Elementos	**Representatividade no Grupo (%)**	**Proporção para 10 laudos**	**Laudos Selecionados**
Cluster 1	157	7,7954319761668	0,93545183714002	1
Cluster 2	57	2,8301886792453	0,339622641509434	0
Cluster 3	117	5,8093346573982	0,697120158887785	1
Cluster 4	0	0	0	0
Cluster 5	1118	55,5114200595829	6,66137040714995	7
Cluster 6	1	0,0496524329692	0,005958291956306	0
Cluster 7	33	1,6385302879841	0,196623634558093	0
Cluster 8	385	19,116186693148	2,29394240317776	2
Cluster 9	0	0	0	0
Cluster 10	146	7,2492552135055	0,869910625620655	1
Total	2014	100%	12	12

Tabela C.10. Laudos de ultrassonografia de abdômen superior a serem extraídos de cada *cluster*

	Elementos	**Representatividade no Grupo (%)**	**Proporção para 10 laudos**	**Laudos Selecionados**
Cluster 1	0	0	0	0
Cluster 2	30	1,1346444780635	0,170196671709531	0
Cluster 3	1760	66,5658093797277	9,98487140695915	10
Cluster 4	478	18,0786686838124	2,71180030257186	3
Cluster 5	55	2,0801815431165	0,312027231467473	0
Cluster 6	111	4,1981845688351	0,629727685325265	1
Cluster 7	60	2,2692889561271	0,340393343419062	0
Cluster 8	91	3,4417549167927	0,516263237518911	1
Cluster 9	53	2,0045385779123	0,300680786686838	0
Cluster 10	6	0,2269288956127	0,034039334341906	0
Total	2644	100%	15	15

Tabela C.11. Laudos de mamografia bilateral a serem extraídos de cada *cluster*

	Elementos	**Representatividade no Grupo (%)**	**Proporção para 10 laudos**	**Laudos Selecionados**
Cluster 1	41	1,1144332699103	0,234030986681163	0
Cluster 2	166	4,5120956781734	0,947540092416417	1
Cluster 3	246	6,6865996194618	1,40418592008698	1
Cluster 4	9	0,2446316933949	0,051372655612938	0
Cluster 5	2615	71,0790975808644	14,9266104919815	15
Cluster 6	349	9,4862734438706	1,99211742321283	2
Cluster 7	100	2,7181299266105	0,570807284588203	1
Cluster 8	102	2,7724925251427	0,582223430279967	1
Cluster 9	30	0,8154389779831	0,171242185376461	0
Cluster 10	21	0,5708072845882	0,119869529763523	0
Total	3679	100%	21	21

Tabela C.12. Laudos de ultrassonografia de próstata via abdominal a serem extraídos de cada *cluster*

	Elementos	**Representatividade no Grupo (%)**	**Proporção para 10 laudos**	**Laudos Selecionados**
Cluster 1	452	10,5533504552883	2,63833761382209	3
Cluster 2	447	10,4366098529068	2,60915246322671	3
Cluster 3	428	9,9929955638571	2,49824889096428	2
Cluster 4	415	9,6894699976652	2,4223674994163	2
Cluster 5	425	9,9229512024282	2,48073780060705	2
Cluster 6	397	9,2692038290918	2,31730095727294	2
Cluster 7	420	9,8062106000467	2,45155265001167	2
Cluster 8	423	9,8762549614756	2,4690637403689	2
Cluster 9	441	10,296521130049	2,57413028251226	3
Cluster 10	435	10,1564324071912	2,53910810179781	3
Total	4283	100%	25	24

Tabela C.13. Laudos de ultrassonografia de tireoide a serem extraídos de cada *cluster*

	Elementos	**Representatividade no Grupo (%)**	**Proporção para 10 laudos**	**Laudos Selecionados**
Cluster 1	586	10,4886343296939	3,461249328799	3
Cluster 2	563	10,0769643816001	3,32539824592805	3
Cluster 3	554	9,9158761410417	3,27223912654376	3
Cluster 4	531	9,5042061929479	3,13638804367281	3
Cluster 5	568	10,1664578485771	3,35493109003043	3
Cluster 6	535	9,5758009665295	3,16001431895472	3
Cluster 7	550	9,8442813674602	3,24861285126186	3
Cluster 8	555	9,9337748344371	3,27814569536424	3
Cluster 9	572	10,2380526221586	3,37855736531233	3
Cluster 10	573	10,255951315554	3,38446393413281	3
Total	5587	100%	33	30

Tabela C.14. Laudos de ultrassonografia com doppler colorido de vasos a serem extraídos de cada *cluster*

	Elementos	**Representatividade no Grupo (%)**	**Proporção para 10 laudos**	**Laudos Selecionados**
Cluster 1	637	10,3560396683466	3,72817428060478	4
Cluster 2	647	10,5186148593725	3,78670134937409	4
Cluster 3	640	10,4048122256544	3,74573240123557	4
Cluster 4	595	9,6732238660381	3,4823605917737	3
Cluster 5	610	9,9170866525768	3,57015119492765	4
Cluster 6	591	9,6081937896277	3,45894976426597	3
Cluster 7	617	10,0308892862949	3,61112014306617	4
Cluster 8	619	10,0634043245001	3,62282555682003	4
Cluster 9	592	9,6244513087303	3,4648024711429	3
Cluster 10	603	9,8032840188587	3,52918224678914	4
Total	6151	100%	36	37

Tabela C.15. Laudos de ultrassonografia de mamas bilateral a serem extraídos de cada *cluster*

	Elementos	**Representatividade no Grupo (%)**	**Proporção para 10 laudos**	**Laudos Selecionados**
Cluster 1	0	0	0	0
Cluster 2	1966	23,6753371868979	11,364161849711	11
Cluster 3	1416	17,0520231213873	8,1849710982659	8
Cluster 4	0	0	0	0
Cluster 5	0	0	0	0
Cluster 6	0	0	0	0
Cluster 7	0	0	0	0
Cluster 8	4922	59,2726396917148	28,4508670520231	28
Cluster 9	0	0	0	0
Cluster 10	0	0	0	0
Total	8304	100%	48	47

Tabela C.16. Laudos de ultrassonografia de abdômen total a serem extraídos de cada *cluster*

	Elementos	**Representatividade no Grupo (%)**	**Proporção para 10 laudos**	**Laudos Selecionados**
Cluster 1	44	0,4844214466586	0,256743366729054	0
Cluster 2	0	0	0	0
Cluster 3	305	3,3579213916107	1,77969833755367	2
Cluster 4	0	0	0	0
Cluster 5	206	2,2679731366289	1,2020257624133	1
Cluster 6	5459	60,101288120665	31,8536827039524	32
Cluster 7	0	0	0	0
Cluster 8	1041	11,460971044809	6,07431465374876	6
Cluster 9	1784	19,6410877463393	10,4097765055598	10
Cluster 10	244	2,6863371132886	1,42375867004294	1
Total	9083	100%	53	52

Tabela C.17. Laudos de ultrassonografia de articulação a serem extraídos de cada *cluster*

	Elementos	**Representatividade no Grupo (%)**	**Proporção para 10 laudos**	**Laudos Selecionados**
Cluster 1	58	0,5322077445403	0,340612956505781	0
Cluster 2	56	0,5138557533492	0,328867682143513	0
Cluster 3	114	1,0460634978895	0,669480638649293	1
Cluster 4	133	1,2204074142045	0,781060745090842	1
Cluster 5	229	2,1013029913746	1,34483391447972	1
Cluster 6	115	1,0552394934851	0,675353275830428	1
Cluster 7	98	0,8992475683612	0,575518443751147	1
Cluster 8	286	2,6243347403193	1,67957423380437	2
Cluster 9	9693	88,9429253073958	56,9234721967333	57
Cluster 10	116	1,0644154890806	0,681225913011562	1
Total	10898	100%	64	65

Tabela C.18. Laudos de ultrassonografia transvaginal a serem extraídos de cada *cluster*

	Elementos	**Representatividade no Grupo (%)**	**Proporção para 10 laudos**	**Laudos Selecionados**
Cluster 1	1582	11,5222141296431	9,10254916241806	9
Cluster 2	1381	10,0582665695557	7,94603058994902	8
Cluster 3	76	0,5535324107793	0,437290604515659	0
Cluster 4	1560	11,3619810633649	8,97596504005827	9
Cluster 5	52	0,3787327021122	0,299198834668609	0
Cluster 6	1379	10,0436999271668	7,93452294246176	8
Cluster 7	1393	10,1456664238893	8,01507647487254	8
Cluster 8	1464	10,6627822286963	8,42359796067007	8
Cluster 9	3367	24,5229424617626	19,3731245447924	19
Cluster 10	1476	10,7501820830299	8,49264384559359	8
Total	13730	100%	79	77