Pontifícia Universidade Católica
do Rio de Janeiro

Suemi Higuchi

Extração automática de informações: uma leitura distante do Dicionário Histórico-Biográfico Brasileiro (DHBB)

Tese de Doutorado

Tese apresentada como requisito parcial para obtenção do título de doutora em Letras/Estudos da Linguagem pelo Programa de Pós-Graduação em Estudos da Linguagem do Departamento de Letras da PUC-Rio

Orientadora: Profa. Maria Cláudia de Freitas

Rio de Janeiro,
abril de 2021

Pontifícia Universidade Católica
do Rio de Janeiro

Suemi Higuchi

Extração automática de informações: uma leitura distante do Dicionário Histórico-Biográfico Brasileiro (DHBB)

Tese apresentada como requisito parcial para obtenção do título de doutora pelo Programa de Pós-Graduação em Estudos da Linguagem da PUC-Rio. Aprovada pela comissão examinadora abaixo:

Profª Maria Cláudia de Freitas
Orientadora
Departamento de Letras – PUC-Rio

Profª Diana de Souza Marques dos Santos
University of Oslo

Prof. Emanoel Cesar Pires de Assis
UEMA

Prof. Carlos Henrique Marcondes de Almeida
UFF

Prof. Leandro Guimarães Marques Alvim
UFRRJ

Rio de Janeiro, abril de 2021

Suemi Higuchi

Graduada em História pela Universidade Federal Fluminense (2000), pós-graduada (MBA) em Gestão de Negócios e Tecnologia da Informação pela FGV (2005), mestre em Ciência da Informação pelo PPGCI/UFF (2012) e doutora em Linguística pelo PPGEL/PUC-Rio (2021) com estágio sanduíche na University of Oslo (2018-2019). É pesquisadora do Centro de Pesquisa e Documentação em História Contemporânea do Brasil da Fundação Getulio Vargas (CPDOC/FGV).

Ficha Catalográfica

Higuchi, Suemi

Extração automática de informações : uma leitura distante do Dicionário Histórico-Biográfico Brasileiro (DHBB) / Suemi Higuchi ; orientadora: Maria Cláudia de Freitas. – 2021.
176 f. ; 30 cm

Tese (doutorado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Letras, 2021.
Inclui bibliografia

1. Letras – Teses. 2. Linguística computacional. 3. Linguística com corpus. 4. Humanidades digitais. 5. Leitura distante. 6. Extração de informações. I. Freitas, Maria Cláudia de. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Letras. III. Título.

CDD: 400

Ao meu pai,

em sua memória.

Agradecimentos

À minha orientadora professora Claudia Freitas pelo suporte, incentivo e enorme generosidade na realização deste trabalho. À professora Diana Santos, pela inestimável ajuda na pesquisa e todo apoio e acolhimento em Oslo. Aos demais membros da banca, pelas generosas e ricas contribuições. À Chiquinha, por sempre me socorrer com gentileza e paciência na secretaria da pós.

À PUC-Rio, CAPES e CPDOC/FGV, pelos auxílios concedidos, sem os quais esta pesquisa não poderia ter sido realizada.

Aos meus amigos de perto e de longe, por serem meu calmante e minha diversão quando precisei. À minha família querida por todo amor, construção e apoio emocional. Àquele que soube me traduzir por inteiro.

A Deus, por tudo.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

"*não amemos de palavra nem de língua,*

mas em ação e em verdade."

(1 João 3:18)

Resumo

Higuchi, Suemi; Freitas, Maria Cláudia (Orientadora). **Extração automática de informações: uma leitura distante do Dicionário Histórico-Biográfico Brasileiro (DHBB)**. Rio de Janeiro, 2021. 176 p. Tese de Doutorado – Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

A pesquisa aplica algumas técnicas de processamento de linguagem natural (PLN) ao domínio da história, tendo como objeto de investigação o Dicionário Histórico-Biográfico Brasileiro (DHBB), obra de estilo enciclopédico concebida pelo Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC) da Fundação Getulio Vargas (FGV). O objetivo foi criar, a partir do DHBB, um corpus anotado para fins de extração automática de informações, relevante para as Humanidades Digitais, capaz de viabilizar 'leituras distantes' da política contemporânea brasileira. O processo completo passa pelas etapas de análise morfossintática do material, identificação de entidades relevantes ao domínio, inclusão de anotação no corpus, definição de relações semânticas de interesse para a pesquisa e mapeamento dos padrões léxico-sintáticos existentes nestas relações. Busca-se com estas etapas preparar os textos para a identificação de estruturas de interesse, isolando as informações relevantes e apresentando-as de forma estruturada. Para testar e avaliar um conjunto de padrões quanto à sua produtividade, foram selecionados como temas de interesse idade de entrada dos biografados na carreira política, formação acadêmica e vínculos familiares. O pressuposto é que utilizando padrões léxico-sintáticos é possível extrair informação de qualidade direcionada ao domínio da História, a partir de um corpus anotado do gênero enciclopédico. Na avaliação dos padrões para a extração do ano de nascimento dos biografados a medida-F foi de 99%, para a extração de relações familiares a medida-F foi de 84% e para informações sobre formação acadêmica o índice de acertos alcançou 99,1%. Essas extrações, por sua vez, permitiram uma leitura distante dos dados do DHBB que nos mostra i) queda da média de idade no que se refere à entrada dos políticos na carreira pública, que passam a se posicionar cada vez mais abaixo dos 40 anos, principalmente os nascidos a partir da década de 1960; ii) declínio acentuado na formação militar, sobretudo para as gerações pós 1920, demonstrando que o treinamento civil estava substituindo o militar enquanto

caminho para atingir cargos políticos importantes; e iii) vínculos familiares na política como um fenômeno que se mantêm ao longo do tempo em índices bastante significativos, muitas vezes representando mais de 50% do total de membros de determinadas categorias. As principais contribuições da tese são: criação de um corpus de gênero enciclopédico anotado e disponibilizado para estudos linguísticos e das humanidades; apresentação de metodologia baseada em uma filosofia de enriquecimento cíclico, em que à medida que se vai obtendo mais informações, elas são adicionadas ao próprio corpus melhorando a extração; e compilação de um conjunto de padrões passível de ser adaptado para quaisquer corpora contendo o mesmo tipo de anotações.

Palavras-chave

Linguística computacional, linguística com corpus, humanidades digitais, leitura distante, extração de informações, Dicionário Histórico-Biográfico Brasileiro

Abstract

Higuchi, Suemi; Freitas, Maria Cláudia (Advisor). **Automatic information extraction: a distant reading of the Brazilian Historical-Biographical Dictionary (DHBB)**. Rio de Janeiro, 2021. 176 p. PhD Thesis – Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

The research applies some natural language processing techniques (NLP) to the domain of history, having as object of investigation the Brazilian Historical-Biographical Dictionary (DHBB), an encyclopedic style work conceived by the Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC) of Fundação Getulio Vargas (FGV). The target is to create, from the DHBB, an annotated corpus for automatic information extraction's purpose, relevant to the Digital Humanities, enabling "distant readings" of Brazilian contemporary political history. The complete process goes through the morphosyntactic analysis of the material, identification of entities relevant to the domain, inclusion of semantic annotation in the corpus, definition of semantic relations of interest and mapping of lexical-syntactic patterns existing in these relations. These steps seek to prepare the texts for the identification of structures of interest, isolating the relevant information and presenting them in a structured way. To test and evaluate a set of textual patterns regarding their productivity in relation to DHBB, some specific topics were selected: age of the politician when entering public life, academic training and family ties. The assumption is that using lexical-syntactic patterns it is possible to extract high quality information from the domain of History, from an annotated corpus of the encyclopedic genre. In the evaluation of the patterns for extraction of the year of birth of the biographees, the F-measure was 99%, for the extraction of family relationships, the F-measure was 84% and for information on academic training, the correctness index reached 99.1%. These extractions, in turn, allowed us to make a distant reading of the data in the DHBB that shows us i) a drop in the average age with regard to the entry of politicians into the public career, who start to position themselves more and more under 40 years of age, mainly those born from the 1960s; ii) sharp decline in military training, especially for the post-1920 generations, demonstrating that civilian training was replacing military training as a way to reach important political positions; and iii)

family ties in politics as a phenomenon that remain over time at very significant rates, often representing more than 50% of the total members of certain categories. The main contributions of the thesis are: creation of an encyclopedic genre corpus annotated and made available for linguistic and humanities studies; presentation of a methodology based on a philosophy of cyclic enrichment, in which, as more information is obtained, they are added to the corpus itself, improving extraction; and compilation of a set of productive patterns that can be adapted for any corpora containing the same type of annotations.

Keywords

Computational linguistics, corpus linguistics, digital humanities, distant reading, information extraction, Brazilian Historical-Biographical Dictionary

Sumário

Lista de figuras

Lista de tabelas

Lista de siglas e abreviaturas

AC/DC – Acesso a corpus/Disponibilização de corpus
ACE – Automatic Content Extraction
CG – Constraint Grammars
CPDOC – Centro de Pesquisa e Documentação de História Contemporânea do Brasil
CQP – Corpus Query Processor
CoNLL – Conference on Computation Natural Language Learning
DHBB – Dicionário Histórico-Biográfico Brasileiro
FGV – Fundação Getulio Vargas
HAREM – Avaliação de Reconhecimento de Entidades Mencionadas
HD s – Humanidades Digitais
HPSG – Head driven frase structure grammars
ICDAR – International Conference on Document Analysis and Recognition
IE (ou EI) – Extração de informação
IberLEF – Iberian Language Evaluation Forum
KWIC – Key word in contexto
ML – Machine Learning
MUC – Message Understanding Conference
MWEs – Multiword expressions
NER – Named Entity Recognition
NIST – National Institute of Standards and Technology
NLP – Natural Language Processing
PLN – Processamento de Linguagem Natural
POS – Part of Speach (tagging)
R – Ambiente e Linguagem de programação
ReRelEM – Reconhecimento de relações entre entidades mencionadas
TAC – Text Analysis Conference
UD – Universal Dependency
VSM – Vector Space Models

1
Introdução

Para mediar a exploração de fontes disponíveis em formato digital, a Linguística Computacional, ou Processamento de Linguagem Natural (PLN), nasce no desafio de fornecer aos sistemas capacidade para reconhecer e extrair informação automaticamente a partir de textos escritos em linguagem humana (Chowdhury, 2003).

Ainda que tenha se constituído como uma ciência fortemente empírica, seu desenvolvimento ao longo dos anos não esteve condicionado apenas aos recursos informáticos disponíveis às épocas, mas também as questões nucleares da Linguística, sensível às formações teórico-filosóficas do campo e as investigações sobre linguagem (Coates-Stephen, 1992; Sparck-Jones, 2001; Hirst, 2009).

Em fins dos anos 1940 e pelas duas décadas seguintes, o campo foi fortemente focado em tradução automática e estudos sobre as regras de linguagem; durante toda a década de 1970 foi influenciado pelos construtos e estímulos da inteligência artificial, principalmente em estudos sobre semântica e representação do conhecimento; nos anos 1980 esteve predominantemente alicerçado por uma teoria gramatical vinculada a lógicas de representação; nos anos 1990 em diante, com o crescimento da Internet e da capacidade de processamento dos computadores e a diminuição gradual da dominância das teorias chomskyanas, se voltou para a resolução de tarefas de análise e extração de informação, validação de sistemas e construção de recursos léxicos, com massivo uso de corpora e estudos estatísticos (Manning & Schutze, 1999; Spark-Jones, 2001; Hockey, 2004), culminando com o avanço na adoção de técnicas de *machine learning* e a aplicação de cálculos complexos no processamento destes corpora (Jurafsky & Martin, 2009).

Hoje, o campo acena com um conjunto de técnicas e ferramentas capazes de realizar tarefas como reconhecimento de entidades nomeadas, identificação de estruturas morfossintáticas e atribuição de informação semântica a porções de texto. Quando falamos especificamente sobre extração automática de informações, tema desta tese, reportamo-nos às atividades que envolvem a extração de entidades e relações a partir de uma coleção de textos. Para a realização desta tarefa, pesquisas vêm se concentrando desde há muito em estudos de padrões léxico-sintáticos para

a produção de regras linguísticas precisas e abrangentes, com ou sem o apoio de algoritmos de aprendizagem de máquina e modelos estatísticos (Hearst, 1992, 1998; Pennacchiotti & Pantel, 2006; Mausam et al., 2012; Makarov, 2018).

O Dicionário Histórico-Biográfico Brasileiro é, originalmente, um compêndio de milhares de verbetes biográficos e temáticos sobre a história política contemporânea do Brasil, que foi convertido em um corpus anotado morfossintaticamente. Partindo de uma abordagem fortemente identificada pelos trabalhos desenvolvidos por Marti Hearst (1992, 1998) –, a informação é extraída a partir de um conjunto de padrões e regras léxico-sintáticos específicos ao domínio. Apesar de sua simplicidade, o método mostra-se bastante preciso (Chiticariu, 2013; Makarov, 2018) e tem a vantagem de ser transparente e prontamente examinável, sendo acionado pela correspondência de estruturas específicas no texto. Por outro lado, a construção manual de padrões pode se mostrar de alto custo (Makarov, 2018), pois não são facilmente transportáveis entre domínios, e qualquer adaptação a um novo domínio requer um esforço humano significativo. A escolha desta estratégia se deve, dentre outros fatores, à previsibilidade dos textos do DHBB – cuja escrita segue uma estrutura bastante padronizada –, ao acionamento direto de regras e léxicos específicos capazes de melhorar a identificação de certos tipos de informação no corpus, e a acessibilidade da abordagem, que não requer domínio computacional complexo na aplicação dos padrões.

1.1 Contexto e motivação

O DHBB reúne mais de 7,5 mil verbetes biográficos e temáticos sobre a história recente do país, e contém informações que vão desde a trajetória de vida, formação e carreira dos indivíduos, até as relações construídas entre os personagens e eventos que o país abrigou (Abreu et al, 2010). A principal motivação para explorar esta obra através de ferramentas da linguística computacional surgiu da necessidade que os pesquisadores têm de buscarem certas informações sem a leitura meticulosa dos verbetes. Em uma consulta realizada junto a alguns pesquisadores e acadêmicos que costumam consultar o DHBB com frequência, perguntamos sobre que questões gostariam de ver respondidas automaticamente caso fosse possível.

Seis desses pesquisadores enviaram suas perguntas[1]. Selecionamos e adaptamos abaixo algumas delas para mostrar a diversidade de temas colocados:

- Quais os políticos que nasceram antes da década de 1960, tiveram formação militar e ocuparam algum cargo no Executivo?
- Como se caracteriza a formação superior dos quadros políticos ao longo das gerações?
- Qual a idade dos ministros do Supremo Tribunal Federal ao serem nomeados?
- Qual o perfil partidário dos ministros do Poder Executivo republicano?
- Quem são os políticos que detêm vínculos familiares com outros políticos? Que vínculos são esses?

Muitas das respostas para estas perguntas se encontram dispersas nos verbetes do DHBB e não estão indexadas em campos de metadados. Um sistema de extração de informações (IE) pode ajudar a vencer este desafio, pois tem como objetivo selecionar e obter informações específicas em meio a grandes volumes de texto.

O foco deste trabalho não é a construção de uma interface que permita responder às perguntas dos pesquisadores, mas sim criar subsídios para tal a partir da exploração de alguns dos principais aspectos que integram a atividade de mineração em corpus, conjugando estudos sobre as estruturas da língua com técnicas e ferramentas computacionais disponíveis. Análises relativas a sintaxe e semântica permeiam a pesquisa.

Temos por um lado o processo de reconhecer as entidades mencionadas e por outro a identificação das relações que as conectam. Perguntas sobre "quem trabalhou onde", "fazendo o quê" e "com quem" se beneficiam de informação de natureza sintática – como a relação entre sujeito, verbo e complemento –, e levam a respostas que envolvem tipos semânticos como "pessoa", "lugar" e "organização", além de relações de caráter institucional, familiar, político etc.

[1] As questões na íntegra enviadas pelos pesquisadores encontram-se no Anexo1.

Analisadores automáticos (*parsers*) são capazes, hoje, de determinar com relativa eficácia a estrutura sintática de um texto por meio da análise de suas palavras constituintes. Debruçar-se sobre as saídas anotadas de um corpus resulta na observação de certos aspectos da linguagem, como distribuição e sintaxe, e de como certos fenômenos se realizam na língua. Todos estes pontos têm forte conexão com temas sensíveis aos estudos da linguagem pois trazem ao palco conceitos circunscritos à palavra e ao sentido.

O presente trabalho busca ampliar o horizonte das relações entre as humanidades e o uso das tecnologias disponíveis. A pesquisa em linguística computacional se apresenta como ponto de partida e de suporte na preparação do DHBB, instrumentalizando os textos para a extração automática de estruturas.

A intenção é que toda a metodologia e processos percorridos possam ser aplicados em outros corpora e outras coleções de documentos de domínios semelhantes, contribuindo com uma área fértil dentro das humanidades digitais. Os resultados têm impacto na ampliação da pesquisa acadêmica nas ciências sociais e humanas sob diversos aspectos, tanto em termos de renovação de métodos quanto de produção de conhecimento.

1.2
Objetivos e hipótese de pesquisa

1. Objetivo Geral: criar, a partir do Dicionário Histórico-Biográfico Brasileiro, um corpus anotado para fins de extração de informação automática baseada nos aspectos morfossintáticos do português e no campo semântico da história política contemporânea brasileira.
2. Objetivos Específicos:
    - Investigar as bases para a extração de informação no domínio da história, identificando os desafios, as contribuições teóricas e metodológicas que melhor endereçam estas bases;
    - Testar a metodologia de extração proposta, apoiada em padrões textuais, usando os seguintes temas: idade de entrada dos biografados na carreira pública, formação acadêmica e vínculos familiares entre políticos.

A principal hipótese da pesquisa é que a utilização de padrões léxico-sintáticos permite extrair informação de qualidade para o domínio da História – e de humanas de um modo geral –, a partir de um corpus anotado de gênero enciclopédico. Ela é colocada à prova na seção 5.4 da metodologia e avaliada na seção 5.5.

1.3 Principais contribuições

Corpora anotados em português – e outros recursos lexicais, de um modo geral – ainda são bastante escassos na comunidade de PLN, sendo toda iniciativa neste sentido bastante válida. Desta forma, as principais contribuições da tese são as seguintes:

- Criação de um corpus de gênero enciclopédico anotado e disponibilizado para estudos linguísticos, históricos e de humanidades digitais;
- Demonstração da eficiência da abordagem adotada, especialmente pela simplicidade e versatilidade na criação das regras, podendo ser lexicalmente incrementada e semanticamente enriquecida a qualquer momento;
- Apresentação de metodologia baseada em uma filosofia de enriquecimento cíclico, em que à medida que se vai obtendo mais informações, elas são adicionadas ao próprio corpus, melhorando a extração;
- Compilação de um conjunto de padrões produtivos para extrair informação específica, podendo ser adaptado para outros corpora do domínio contendo o mesmo tipo de anotações;
- Uma contribuição não esperada, porém importante, foi a identificação de dados incorretos ou incompletos nos verbetes, que foram surgindo ao longo do processo, permitindo a própria melhoria do DHBB.

1.4
Organização da tese

A tese está organizada da seguinte forma: nesta introdução foram apresentados o contexto, motivação e questões que orientaram a pesquisa, além dos objetivos e principais contribuições a serem alcançados. No segundo capítulo, investigam-se os aportes teóricos e metodológicos recorrentes nas áreas relevantes para o trabalho. No terceiro capítulo, o DHBB é apresentado na sua importância como referência para estudos sobre a história política contemporânea do Brasil, e nas possibilidades de pesquisa como corpus anotado. No capítulo seguinte, de metodologia, são apresentados os processos pelos quais a tarefa de extração de informações é conduzida. O quinto capítulo é dedicado às análises dos resultados, através de pequenas 'leituras distantes' feitas com os dados obtidos. No sexto e último capítulo, traçamos as considerações finais e possibilidades de trabalhos futuros.

2
Aspectos teórico-metodológicos

Este capítulo visita os aportes teóricos e metodológicos recorrentes nas áreas de pesquisa relevantes para este trabalho. Em suma, a tese conecta conceitos, abordagens e ferramentas das seguintes disciplinas:

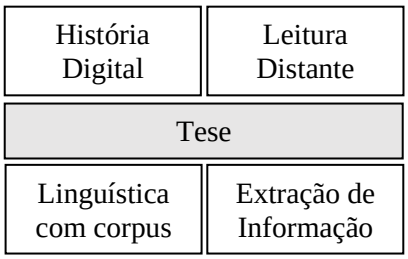

Figura 1 - Principais áreas de pesquisa da tese

O domínio de aplicação é a história, em particular o ‘fazer história’ na era digital. Estamos vivendo um momento em que repensar a disciplina e os métodos de pesquisa até então utilizados é quase uma obrigação diante das rápidas mudanças que a tecnologia impõe. Seria menos um movimento para transformar a história, e mais para garantir que os historiadores e outros humanistas continuem – ou ainda, melhorem – a sua capacidade de analisar os registros históricos neste contexto em que quantidade e velocidade muitas vezes impactam nossas descobertas.

Uma das aplicações mais promissoras tem sido a técnica conhecida nas humanidades digitais como “leitura à distância” ou “leitura distante”. A leitura distante é uma inversão deliberada do termo mais familiar “leitura atenta”, que significa um exame cuidadoso e minucioso das particularidades de um texto. Em contraste, a leitura distante envolve o uso de automação para fazer generalizações sobre vastos corpora. Enquanto um texto é para ser lido horizontalmente prestando-se atenção às cláusulas, sentenças e parágrafos, um corpus é varrido verticalmente, buscando-se por padrões presentes em contexto (Moretti, 2000).

O desenvolvimento de ferramentas apoiadas em corpora abrange áreas tão distintas como ensino e aprendizado de línguas, análise do discurso, linguística forense, tradução, pragmática, tecnologia da fala, sociolinguística, entre outras. As técnicas e abordagens variam e dependem da aplicação. No caso desta tese, é a

extração de informações (IE) que norteará o trabalho, conjugando a descoberta de padrões para a obtenção automática de estruturas a partir do corpus DHBB.

2.1
História digital

De que forma a prática de pesquisa no domínio da história tem sido modificada com o uso de ferramentas digitais? Que desafios precisam ser enfrentados e que oportunidades se abrem neste cenário potencialmente inovador? Reflexões acerca destas questões vem sendo progressivamente endereçadas por autores que tentam delinear as características dessas transformações e suas limitações (Seefeldt & Thomas, 2009; Lucchesi, 2014; Brasil & Nascimento, 2020).

História digital é um termo pouco demarcado em termos epistemológicos, mas pode ser compreendido de forma ampla como uma abordagem para examinar e representar o passado que funciona em conjunto com as novas tecnologias de comunicação mediadas pelo computador, a rede de Internet e os sistemas de software (Seefeldt e Thomas, 2009). O cenário envolve a pesquisa em si, o modo como passamos a reunir e analisar as fontes no meio digital e as mudanças na difusão do conhecimento adquirido, através das salas de aula, das redes sociais e das publicações online.

Alguns autores mostram-se céticos ou não veem muito sentido em tentar rotular uma disciplina onde em breve – como todas as outras –, o digital estará tão incorporado nas suas práticas que será parte inerente dela (Frisch, 2008; Alves, 2017). Em uma entrevista publicada em 2017, Alves argumenta que:

> *O termo História Digital foi em grande medida ultrapassado ou confundido com as Humanidades Digitais, mas, ao mesmo tempo, hoje, o digital já envolve muito do que os historiadores fazem, mesmo que nem todos se apercebam ou sequer o pretendam valorizar. Nessa perspectiva, tal como me parece que daqui a uns anos falar de Humanidades será essencialmente falar de Humanidades Digitais, hoje em dia falar de História já é muito falar de História Digital* (Alves, 2017).

Assim, o fato de a tecnologia estar tão integrada em todas as atividades cotidianas e profissionais propicia tal cenário de naturalização ao mundo digital.

2.1.1
O domínio da história e as ferramentas digitais

A despeito das discussões sobre demarcações do campo, é inegável que com as ferramentas digitais é possível mostrar muito mais das operações da história, ou seja, as histórias embutidas nos dados primários e as negociações e decisões que levam às estruturas, ideias e formas das narrativas. Tanaka (2013) argumenta que na escrita da história, tradicionalmente, a pesquisa é fundamentada no gerenciamento de uma infinidade de informações e dados que resultam em uma narrativa unitária mais ou menos direta, onde os estudos são limitados ao comprimento de um livro ou artigo, e decisões são tomadas sobre pontos de vista conflitantes, muitas vezes com a omissão de contradições. Mas no meio digital tem-se a oportunidade do uso de ferramentas que facilitam narrativas mais complexas e dimensionadas sobre o passado. O papel do historiador muda de especialista que domina (e protege seu) conhecimento sobre uma área muito específica e cada vez mais estreita, para o de um organizador qualificado e confiável de uma miríade de dados que ajudam a compreender a experiência humana (Tanaka, 2013).

Pelo lado empírico, talvez o maior desafio esteja no impacto da superabundância de dados, que altera significativamente a forma como o historiador passa a elaborar sua visão do passado. Essa é a mesma percepção de Daniel Alves, quando observa que "habituado à escassez de informação, a dados lacunares e dispersos, o problema daqui para a frente poderá ser o da seleção e avaliação da pertinência de um grande volume de dados. E isso implicará uma crítica face ao digital, quer a fonte, quer a ferramenta" (Alves, 2017).

Para Cohen há uma grande chance de o historiador "analógico" vir a ter seu trabalho de pesquisa depreciado a depender do contexto: porque muitas vezes os massivos conjuntos de dados à disposição podem fazer parecer frágeis hipóteses baseadas em um número limitado de exemplos, e até mesmo anedóticos os argumentos históricos que tentam ser abrangentes, mas ignoram deliberadamente contra-exemplos localizados nesses conjuntos digitais (Cohen, 2008:456).

Por outro lado, é preciso refletir sobre o problema da especificidade do uso das novas tecnologias e recursos no campo historiográfico. Apesar de ser possível fazer uso de métodos estatísticos complexos para produzir conhecimento, Gibbs e

Owens (2013), argumentam que, ao contrário do campo das ciências exatas, na história o rigor matemático não é essencial para um uso eficaz dos dados, que poderiam ser abordados, por exemplo, de modo exploratório. Definindo dados como as informações processáveis por computador, os autores alertam para o cuidado de não se confundir dados e evidências, ainda que os dados possam, em alguma circunstância específica, serem utilizados como evidência para algum argumento histórico.

Em outras palavras, podemos dizer que os *dados* são o resultado de um processo de obtenção de dados – tal como os dados gerados a partir do DHBB ou a compilação, por exemplo, dos anais do Congresso Nacional – e não uma representação direta do registro histórico.

2.1.2
O ofício do historiador

Por muitos anos o que se convencionou chamar de história quantitativa constituiu um capítulo particularmente importante no campo historiográfico, principalmente em fins do século XIX e meados do XX. Aos textos foram agregados números e estatísticas gerados sob a preponderância da história econômica e do modelo estrutural vigente, em especial após 1929 com a quebra da bolsa de valores americana e seus múltiplos desdobramentos (Burke, 1997). Quase meio século depois, o célebre historiador Emmanuel Le Roy Ladurie declarava em um artigo de 1968 que "O historiador do futuro será programador ou não será" (Le Roy Ladurie, 2011). Era a visão de um pregador da história serial, que defendia a disposição de fatos em séries temporais de unidades homogêneas e comparáveis, numa proposta de análise do passado a partir da decomposição da realidade segundo critérios quantitativos previamente estabelecidos.

No entanto, o historiador não só não se tornou programador, como também não precisou se tornar programador para fazer uso da informática. Certamente isso se deve em muito à própria transformação da ideia de informática, que inicialmente era associada ao cálculo e à tabulação de dados, mas se tornou mais acessível e amigável através de aplicativos que não exigem necessariamente algum conhecimento de programação.

Hoje, quase todas as problemáticas tradicionais do ofício de historiador – da delimitação de uma hipótese de pesquisa à descoberta, acesso e gestão das fontes, da construção dos fundamentos narrativos à comunicação dos resultados de pesquisa – agora passam, em parte ou no todo, pela tela do computador (Noiret, 2015). Sem dúvida, há uma renovação em termos de metodologia e de produção do conhecimento na área, mas nenhuma dessas ferramentas substitui o trabalho minucioso, dedicado e paciente do historiador. Como Brasil e Nascimento apontam, elas "potencializam nossas habilidades analíticas, possibilitam que novas perguntas sejam formuladas, e novas respostas, atingidas. Mas com muito trabalho e rigor diante do computador" (Brasil & Nascimento, 2020:216).

Fortes e Alvim (2020), ao discutirem sobre a natureza do ofício do historiador e do conhecimento que é produzido diante do impacto da revolução digital, apontam como importante ponto de reflexão o potencial da massiva ampliação do universo de fontes acessíveis e das ferramentas tecnológicas para a produção de análises de qualidade superior no que diz respeito à "inteligibilidade do processo histórico" (Fortes & Alvim, 2020:211). Torna-se mais que necessário pensar em como, quando e de que forma podemos nos beneficiar dos recursos computacionais sem que deixemos de lado a importância da tarefa de pesquisa, crítica e contexto presentes nos trabalhos historiográficos.

2.1.3
Conexões interdisciplinares

O campo das humanidades digitais (HDs) tem sido particularmente fértil para pesquisas localizadas no quintal da história. Segundo Gold e Klein (2016), se em um primeiro momento a caracterização do campo estava na curadoria de acervos digitais e nas análises quantitativas sobre textos – das quais historiadores muito se apropriam –, não demorou muito para incorporar um leque muito mais amplo de métodos, práticas e estudos, tais como visualizações através de largos conjuntos de imagens, modelagem 3D de artefatos históricos, georreferenciamento sobre mapas virtuais, análises de publicações em redes, entre outros.

Na percepção de Daniel Alves, os historiadores mais comprometidos com as HDs cada vez menos publicam de forma isolada e colaboram em projetos de investigação cada vez mais multidisciplinares (Alves, 2017). Um fator que

certamente contribui para tal seria a tradicional falta de habilidade com programação e pensamento algorítmico desses pesquisadores, que os levam a estabelecer colaborações com parceiros mais tecnológicos, como cientistas da computação e engenheiros (Crymble, 2015; Bonfiogli, 2015). O próprio acesso às fontes e recursos digitais, muitas vezes em formatos múltiplos e desconhecidos, também exigirá conexões interdisciplinares.

No entanto, apesar destas interações levarem a um bom número de projetos de pesquisa conjunta com sucesso, também é perceptível que estas colaborações podem ser difíceis de conduzir (Crymble, 2015), pois diferentes formações, abordagens e expectativas precisam ser constantemente focados em um objetivo comum. Não à toa, durante a última década esta lacuna de conhecimento em métodos computacionais tem levado muitos humanistas digitais a preferirem estudos exploratórios realizados a partir de *toolboxes* – por exemplo, o Mallet, que não demanda conhecimento em programação para obter uma modelagem de tópicos –, ao invés de investigações apoiadas em hipóteses quantitativas (Bonfiglioli, 2015).

Ainda assim, projetos complexos envolvendo conhecimento intensivo de computação ou análise estatística, têm sido cada vez mais tocados em conjunto. Ao refletir sobre os papéis exercidos pelos participantes desses projetos multidisciplinares, Crymble escreve em seu artigo que historiadores estão cada vez mais se tornando clientes dos departamentos de ciência da computação, numa clara alusão ao fato de que são eles, os historiadores, que costumam procurá-los com as ideias iniciais e com os problemas que precisam resolver, lançam os desafios de maneira interessante e submetem as propostas para os órgãos de fomentos da área de humanidades e quase nunca em áreas como engenharia ou ciência da computação (Crymble, 2015). Ou seja, na visão do autor, projetos assim resultam particularmente caras para a pesquisa em humanidades.

A despeito dessa visão, é inegável que todas as áreas se beneficiam dessas conexões, independente da parte que lhes cabem na divisão das tarefas e dos custos envolvidos. Métodos e técnicas computacionais só têm razão de ser quando existem problemas reais para serem resolvidos com tais recursos e habilidades, e vice-versa.

2.2
Leitura distante

O imbricamento entre as práticas tradicionais de registro do conhecimento e as novas tecnologias é a marca indelével do movimento das Humanidades Digitais (HDs). Elas incorporam os métodos e questões desenvolvidos pelas ciências humanas e sociais, ao mesmo tempo que mobilizam as ferramentas e perspectivas únicas abertas pela tecnologia digital (Schnapp et al, 2009). Na área mais intimamente ligada à linguagem e literatura, temos observado de forma crescente o uso de métodos quantitativos de mineração de texto para analisar grandes coleções digitais (Moretti, 2013; Bonfiglioli & Nanni, 2015; Wiedemann & Nielker, 2017).

Franco Moretti identificou essa prática com o conceito de leitura distante ou leitura à distância (*distant reading*), quando o distanciamento "é uma condição de conhecimento: permite que você se concentre em unidades que são muito menores ou muito maiores do que o texto: dispositivos, temas, tropos - ou gêneros e sistemas. E se, entre o muito pequeno e o muito grande, o próprio texto desaparece, bem, é um daqueles casos em que se pode dizer com razão, Menos é mais."[2] (Moretti, 2000:57).

2.2.1
Novas escalas de observação

Sem dúvida, os computadores tornaram-se aliados valiosos, proporcionando maneiras inéditas de leitura e permitindo *insights* sobre grandes corpora. Embora as máquinas não possam ler e entender um romance da maneira que as pessoas podem, elas são muito boas em procurar informações específicas e identificar padrões, tanto linguísticos como estruturais, que não seriam visíveis no ato da leitura a 'olho nu' (Jockers, 2012; Freitas, 2015).

Não é um método novo: a abordagem interpretativa de textos apoiada em distribuições e frequências, por exemplo, existe há séculos (Santos et al, 2020:281). Entretanto, a evolução da tecnologia abre novas possibilidades nesse lidar com as fontes, permitindo o emprego de outras escalas de observação. As tarefas de

[2] No original: "Distant reading: where distance […] is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems. And if, between the very small and the very large, the text itself disappears, well, it is one of those cases when one can justifiably say, Less is more."

quantificação ganham nova dimensão. Números e estatísticas não são apenas dados importados de outras fontes, mas podem emergir do próprio texto, em uma relação direta com a linguagem. É possível, por exemplo, medir e comparar o comprimento de sentenças, observar padrões sintáticos e quantificar as variações lexicais do texto, sintetizando dados para outras investigações (Santos, 2014).

Essa perfeita combinação entre dados quantitativos – para conclusões qualitativas – e dados qualitativos – para conclusões quantitativas – tem ajudado especialmente na análise de grandes volumes de fontes textuais (Santos, 2014: 198), hoje cada vez mais acessíveis digitalmente na forma de corpus. Ao se referir aos estudos da história literária, Moretti (2013:66) aponta que os estudiosos até então se concentravam em uma seleção de poucas centenas de textos, devolvendo como quadro geral aquilo que se observava em uma fatia estreita e distorcida da literatura. O caminho para uma história literária mais racional seria substituir a leitura atenta por modelos abstratos emprestados das ciências.

Inspirado pela micro e macroeconomia, Matthew Jockers utiliza o termo *macroanálise* para descrever os métodos estatísticos aplicados nesse tipo de análise em textos (Jockers, 2013). Segundo o autor, a nova abordagem irrompe para o mundo textual como ele existe hoje, em forma digital e em larga escala. Obviamente, coisas importantes podem escapar ao macro, no entanto a história literária não pode ser definida a partir da leitura aproximada de uns poucos autores canônicos. As duas escalas de observação devem e precisam coexistir considerando estas novas formas de recolha de evidências.

2.2.2
Promessas e desconfianças

Inúmeros acadêmicos abraçaram as ideias de Moretti, apontando que métodos computacionais poderiam representar uma alternativa sólida às abordagens hermenêuticas tradicionais, não apenas nos estudos literários, mas também na pesquisa histórica que precisa lidar com grandes volumes de fontes em formato digital (Jockers, 2013; Bonfiglioli, 2015).

Embora seja apenas uma das inúmeras disciplinas (ou metodologias) em HDs, a leitura distante tem sido vista como um bom exemplo das promessas, e sobretudo, desconfianças, do campo. Porque para muitos, ela não produziu

descobertas interessantes e as que hoje produz não podem ser consideradas totalmente confiáveis (Araújo, 2016; Hammond, 2017).

Do rol das críticas recebidas, uma delas traz a ideia de que os métodos computacionais parecem mover-se na direção de tornar o trabalho do humanista irrelevante para a produção de conhecimento original ou inédito, que poderiam ser obtidos bastando apenas o emprego de estatísticas e aprendizado de máquina. Além disso, reduz todos os aspectos dos estudos à busca pela quantificação das características, aspectos e evidências presentes no corpus (Buonfiglioli, 2015:4), como se fosse este o objetivo maior da pesquisa.

Outro autor, Adam Kirsch (2014) diz que os humanistas digitais costumam alardear sobre a grande capacidade de processamento que suas pesquisas passaram a oferecer, mas são incapazes de realizar análises inéditas propriamente ditas. Essa visão pessimista destaca que a adoção indiscriminada da computação pelos acadêmicos muitas vezes não pressupõe a compreensão exata acerca do que acontece aos dados (Ribeiro et al, 2020). Ao usarem as ferramentas, os pesquisadores depositam uma confiança cega nos algoritmos que produzem os resultados, e acabam por não os examinar com o cuidado devido (Dobson, 2015).

2.2.3
Abordagens complementares

Partindo da premissa de que ferramentas digitais podem resultar em pesquisas consistentes e beneficiar estudos das humanidades, é importante destacar a simbiose que deve existir entre leitura distante e leitura atenta.

A experimentação de novas ferramentas e a incorporação do componente computacional às investigações históricas deixou evidente a importância de se ter um olhar qualitativo sobre esses dados quantitativos, evitando-se cair numa espécie de ‘fetichismo’ dos recursos computacionais que imaginasse que eles funcionariam por si só, revelando verdades ocultas no corpus analisado (Castro et al, 2021). Muito pelo contrário, fica evidente que a visão de especialistas é fundamental, desde a elaboração das questões iniciais, passando pela tomada de decisões mais técnicas até a análise intelectual dos resultados.

Em geral, a primeira etapa de um trabalho de processamento de texto deve lidar com a formalização da tarefa de pesquisa, com a adaptação da técnica

computacional escolhida e com a definição das variáveis em uma representação que os algoritmos possam entender. Um outro momento do processo consiste em abstrair conclusões úteis para a pesquisa a partir das saídas geradas pelas análises. Embora um método computacional possa capturar relacionamentos adicionais no corpus, ainda é função do humanista identificar os corretos e, em seguida, validá-los diretamente como novos *insights*, usá-los para tirar novas conclusões ou simplesmente descartá-los. Nesta etapa o especialista deve entender se há causalidade por trás das correlações ou decidir voltar à primeira etapa e ajustar o modelo ou as variáveis, observando os resultados atuais (Bonfiglioli, 2015). Mais uma vez, um forte conhecimento do domínio é claramente importante nesta etapa.

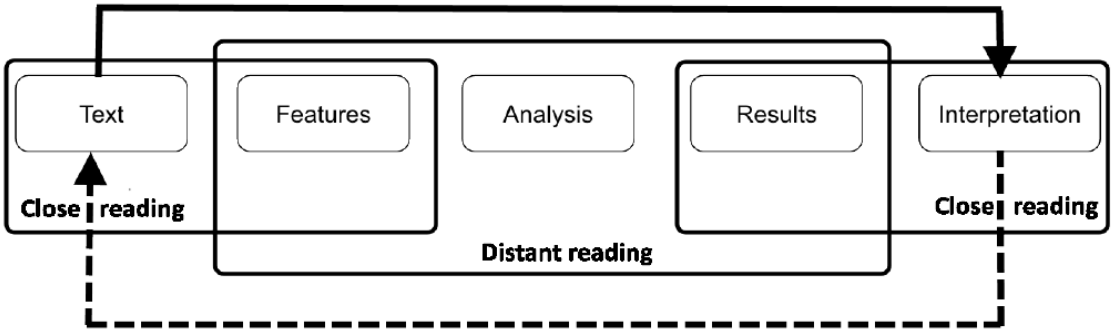

Figura 2 - modelo de interação entre leitura distante e leitura aproximada (extraído de Bonfiglioli, 2015)

O esquema da figura 2 traz uma proposta de modelo de interação entre leitura distante e leitura aproximada para pesquisas em HDs, onde a complementaridade entre ambas as abordagens é o conceito chave.

Pesquisas recentes no campo do aprendizado de máquina têm se debruçado no chamado *deep learning* ou aprendizado profundo, uma família de algoritmos que, dentre outras coisas, visa aprender automaticamente o que seriam boas *features* para extrair informação útil em sistemas IE, substituindo o trabalho feito manualmente pelo especialista e, por consequência, "tornando as máquinas independentes do conhecimento humano" (Najafabadi, 2015:4). Trabalhos com corpora que incorporem técnicas de aprendizado profundo se converteriam então em algo como "leitura distante profunda" (Bonfiglioli, 2015:9).

2.3
Linguística com corpus

Um corpus é uma coleção de textos na forma eletrônica, selecionada de acordo com critérios externos para representar, tanto quanto possível, uma língua ou variedade de uma língua como recurso de dados para pesquisa linguística (Sinclair, 2005). Nesse sentido, estudos com corpus fazem uso de uma abordagem empirista e tem como central a noção de linguagem enquanto sistema probabilístico, sendo possível evidenciar e quantificar regularidades ou padrões no seu uso (Manning & Schutze, 1999).

Por outro lado, existe uma correlação entre os traços linguísticos e os contextos situacionais que não podemos deixar de lado. Nesta seção abordaremos algumas importantes questões relacionadas à linguística com corpus, tanto em suas vertentes mais teóricas como práticas.

2.3.1
Linguagem e sentido

Para o historiador britânico E. H. Carr (1892-1982), o processo de reconstituição que governa a seleção e interpretação dos fatos são ações derivadas da mente de quem escreve, imbuído das suas próprias experiências (Carr, 1978), o que supostamente o impede de ser totalmente neutro na escolha dos termos mais adequados para escrever a História. Por outro lado, a leitura e interpretação destes registros são fenômenos diretamente ligados à capacidade de abstração conceitual – humana ou computacional – que possibilita circunscrever palavras aos seus significados. Grosso modo, significação associa um objeto, um ser, uma coisa, uma noção ou um acontecimento a um signo capaz de evocá-los, e está vinculada, entre outras coisas, às representações que fazemos dos conceitos e das construções nos quais estes participam.

Cara para a Linguística, a questão do signo já foi muito debatida, e ainda o é, por vários autores seminais, como Ferdinand de Saussure e Jacques Lacan. Para o primeiro o significado é escorregadio: “se um objeto pudesse, onde quer que seja, ser o termo sobre o qual é fixado o signo, a linguística deixaria instantaneamente de ser o que ela é, do topo até a base (Saussure, 2002:197). Para o segundo, mais que isso, a linguagem entendida a partir do signo fica aprisionada nessa relação

necessária entre significado e significante, provocando, "a ilusão de que o significante responde à função de representar o significado" (Lacan, 1998:501).

Para além do sentido isolado das expressões linguísticas, Wittgenstein (1979) observa que as funções práticas da linguagem precisam ser levadas em consideração. Segundo o autor, a linguagem não é uma coisa morta em que cada palavra representa algo de uma vez por todas, mas ao contrário, precisa considerar as influências que participam do processo de construção do pensamento socialmente instituído, e, por conseguinte, do significado das coisas. Nesse processo, não haveria como se instituir verdades, apenas construir certezas situacionais (Gracioso e Saldanha, 200); é somente no contexto de uma sentença que a palavra tem significado (Rorty, 1997).

No âmbito da computabilidade, Almeida e Souza (2011:36) afirmam que "a especificação semântica não corresponde precisamente ao significado dos termos, mas sim ao significado de sentenças de acordo com uma *função interpretação* sobre expressões sintaticamente bem formadas de um dado modelo de mundo". Assim, "saber o significado de uma sentença equivale a conhecer suas condições-verdade, o que não é o mesmo que saber o seu valor-verdade, ou seja, se o fato é verdadeiro ou não" (Almeida e Souza, 2011:36).

No contorno geral de um sistema de processamento da linguagem natural a ambiguidade do sentido da palavra se caracteriza como um importante estímulo desde os anos 1950s (Ide & Véronis, 1998). Soluções são buscadas apoiadas em técnicas que combinam recursos lexicais genéricos, como dicionários, tesauros e redes semânticas como as wordnets, além de classificadores estatísticos supervisionados ou não (Manning & Schutze, 1999; Nadeau et al, 2006; Rademaker et al, 2017).

Quando se trata de reconhecimento de entidades mencionadas, Santos (2007) defende uma forma de fazer PLN mais dirigida pelo contexto e menos pelo léxico. As palavras somente podem ser consideradas unidades de sentido quando se encontram em contexto, operacionalizadas nas citações do corpus. Ou seja, os sentidos das palavras somente serão definidos em relação a um conjunto de interesses; dessa forma, o conjunto de sentidos definido pelo dicionário pode ou

não corresponder ao conjunto que é relevante para uma determinada aplicação de PLN (Manning & Schutze, 1999:230).

Quando uma pessoa compreende uma sentença que abriga uma palavra ambígua, esta compreensão é construída tendo por base apenas um dos significados. Então, como parte do processo de compreensão da linguagem humana, o significado apropriado é escolhido a partir de um leque de possibilidades. Para Hirst (2008) o significado de um texto é a soma dos significados das suas sentenças – passíveis de serem processadas – e constante para todos os usuários competentes daquela língua. Para Kilgarriff (2003), o conjunto dos sentidos das palavras para uma língua é dependente da tarefa e é o corpus que dita tais sentidos, não os dicionários.

Tais incursões favorecem reflexões teóricas e metodológicas importantes e minimizam os desassossegos movidos por um campo que não é regido por leis universais e invariáveis, mas também interpretativas, e ajudam a lidar de forma mais adequada com as informações disponíveis nestas narrativas. Na sua *Introdução à Análise Estrutural da Narrativa*, Roland Barthes (1976) afirma que:

> *...diante da infinidade de narrativas, da multiplicidade de pontos de vista pelos quais se podem abordá-las (histórico, psicológico, sociológico, etnológico, estético, etc.), o analista encontra-se na mesma situação que Saussure, posto diante do heteróclito da linguagem e procurando retirar da anarquia aparente das mensagens um princípio de classificação e um foco de descrição [...] O discurso tem suas unidades, suas regras, sua 'gramática'* (Barthes, 1976:20).

Para o autor, todas as unidades possuem um significado dentro da narrativa, e estas unidades se correlacionam de várias formas, dando sentido ao todo no final. Já Fairclough propõe uma análise do discurso que reúna a análise linguística e a teoria social como método para revelar conexões e causas ocultas nos textos, levando em consideração o contexto ao qual estão ligadas (Fairclough, 2008).

Não obstante todas estas manifestações, a escolha de como um conceito é expresso pode revelar informação sobre as ideologias contidas em uma narrativa ou a relação entre participantes da conversa (Wilson e Thomas, 1997). Uma ilustração

disso é o texto do verbete sobre o movimento deflagrado em 31 de março de 1964, incluído no DHBB. Nele, o evento é denotado de duas formas canônicas. Defensores e participantes referem-se a ele como "Revolução de 1964", por considerarem que o seu objetivo era produzir uma reformulação completa na vida política do país, eliminando a corrupção e os mecanismos de poder que estariam sendo utilizados para favorecer a subversão comunista no Brasil; seus opositores e adversários, no entanto, definem-no como "Golpe de 1964", por tratar-se da deposição de João Goulart, um presidente que foi eleito legitimamente pelo povo[3]. O que há é uma situação em que precisamos ser capazes de identificar tanto os termos que se relacionam um ao outro semanticamente quanto os sentidos das palavras em determinados contextos, para que, mais do que expandir a busca, não venhamos a recuperar informação que não nos seja útil (Garside et al, 1997). O ideal é que os sistemas reconheçam que "golpe de 64", "movimento de 64", "regime militar" e "ditadura" são conceitos conexos que se relacionam para mapear um mesmo campo semântico, neste caso, um período da história política brasileira.

O conceito de gramaticalidade, tão caro aos estruturalistas, se preocupa fundamentalmente com a boa formação das sentenças a partir de propriedades finitas da língua. Sendo verdade ou não que "todas as gramáticas vazam" (Sapir, 1949), hoje fenômenos linguísticos complexos, como metáforas, colocações, vaguezas e ambiguidades estão mais passíveis de serem explicados através de modelos estatísticos que medem a distribuição das palavras e expressões nos contextos em que aparecem. Já dizia Wittgenstein (1979), que o significado de uma palavra é definido pelas circunstâncias de seu uso.

Por fim, para Helena Martins (2005) o propósito da linguagem "se manifesta patentemente em sua própria estrutura; caso suas partes (os nomes) não estejam em conformidade com o seu propósito, a linguagem não funciona; para utilizá-la corretamente, precisamos conhecer e respeitar sua arquitetura e seu propósito". (Martins, 2005:459).

[3] A discussão sobre o tema é apresentada no próprio verbete "Golpe de 1964", no DHBB (Abreu et al, 2010).

2.3.2
Pesquisa e consolidação

Em 1957, Noam Chomsky publicou *Syntactic Structures*, que veio a se tornar um divisor de águas na linguística do século XX. A partir daí, desenvolve o conceito de uma gramática gerativa, que se distanciava do estruturalismo e do behaviorismo das décadas anteriores, traçando uma distinção fundamental entre o conhecimento que uma pessoa tem das regras de uma língua – 'competência' – e o uso efetivo desta língua em situações reais – 'performance' (Weedwood, 2002:132; Marcondes, 2009). A linguística, para Chomsky, deveria ocupar-se com o estudo da competência e não do desempenho, em uma clara crítica aos linguistas que buscavam sustentar seus trabalhos baseados no uso de amostras ou corpora. Para ele, tais amostras seriam inadequadas porque representavam apenas uma fração ínfima dos enunciados que é possível dizer numa língua, ou seja, o importante mesmo era focar na descrição das regras que governam a estrutura da competência (Sardinha, 2000; Weedwood, 2002: 133).

Em oposição a Chomsky, o linguista de tradição empirista Michael Halliday, acena a partir dos anos 1960, com uma outra abordagem que ficou conhecida como linguística sistêmica (Weedwood, 2002:137), em que a linguagem é vista enquanto sistema probabilístico dependente dos contextos sociais de uso pelos falantes. Esta visão significa dar primazia aos dados provenientes da observação da linguagem, geralmente reunidos sob a forma de um corpus. Assim, Sardinha destaca duas considerações importantes da abordagem defendida por Halliday:

> *A primeira é a importância primordial de um corpus como fonte de informação, pois ele registra a linguagem natural realmente utilizada por falantes e escritores da língua em situações reais. A segunda é a não-trivialidade da investigação da frequência de ocorrência de traços linguísticos de várias ordens (lexicais, sintáticos, semânticos, discursivos etc.), pois é através do conhecimento da frequência atestada que se pode estimar a probabilidade teórica.* (Sardinha, 2000).

Dessa forma, teoricamente falando, a utilização de corpus nos estudos linguísticos representa um deslocamento das premissas originais chomskianas, onde o foco passa a ser eminentemente na performance ao invés da competência. O objetivo do linguista seria mais o de descrever o uso da linguagem do que identificar universais linguísticos, e o elemento quantitativo (frequência de ocorrências) considerado relevante e, dependendo da abordagem, usado para determinar as categorias da descrição da língua (Bonelli, 2010).

O linguista britânico Geoffrey Sampson, ao contrário de Chomsky, via a Linguística como uma disciplina mais empírica que apriorística (Sampson, 2001:80), tendendo à valorização do método científico, calcado na observação e realização de testes capazes de levar a padrões, que por sua vez subsidiam hipóteses gerais e levam a explicações. Para ele, uma das formas de se estudar e/ou descrever uma língua vem por meio da observação de grandes quantidades de texto, ou seja, a linguagem é algo concreto e tangível, e sendo material, é possível aplicar-lhe técnicas empíricas (Sampson, 2001:1).

Ferramentas estatísticas e recursos sofisticados para exploração de corpora anotados permitem estudos que confirmam ou não hipóteses linguísticas preestabelecidas (Santos, 2008). Em (Santos et al, 2015), os autores descrevem o trabalho da Gramateca[4] e afirmam que a intenção do recurso é contribuir com a metodologia científica no campo da linguística, isto é, não só permitir a repetição de uma experiência – que é uma das propriedades exigidas à metodologia científica –, mas também partilhar diferenças de interpretação de um mesmo corpus: "enquanto nas ciências naturais se espera que a mesma experiência leve aos mesmos resultados, nas ciências humanas é não só esperável, mas provável, que haja diferenças na interpretação quando algo é repetido por outros pesquisadores" (Santos, 2015:13).

Descobertas interessantes podem ser encontradas quando o pesquisador se debruça sobre dados reais. Um estudo estatístico sobre estruturas de oração realizado com o corpus Lancaster-Leeds Treebank demonstrou que, ao contrário do que os livros de linguística pregavam para o inglês, não era verdade que sentenças

[4] Gramateca: um ambiente para fazer uma gramática da língua portuguesa baseada em corpos. Disponível em: http://www.linguateca.pt/Gramateca/

do tipo “sujeito – verbo intransitivo” são as construções mais encontradas ao lado de “sujeito – verbo transitivo – objeto”. Apesar de exemplos como *the lazy child slept* serem a forma básica daquele primeiro tipo, o que se vê na realidade é que, na falta de um objeto seguindo o verbo, quase sempre há algum constituinte ocupando este espaço, como um elemento adverbial ou outro complemento qualquer (Sampson, 2001:92). É um engano achar que exemplos criados nas gramáticas tradicionais são o espelho da vida real, o que definitivamente não é.

Pennycook (2004) recorda que um dos gestos iniciais para a consolidação da Linguística no início do século XX foi o estabelecimento de uma divisão entre linguística interna – a se dedicar unicamente aos estudos da estrutura da língua – e externa – onde a língua é examinada em sua relação com fenômenos sociais, geográficos, culturais e outros, com a valorização da primeira em detrimento da segunda. No entanto, ele ressalta que não é possível entender os mecanismos internos da língua separadamente dos seus usuários e contextos sociais de uso (Pennycook, 2004:40).

Em suma, a pesquisa linguística baseada em corpus compreende dimensões tanto teóricas quanto metodológicas, e diferentes maneiras de entender e estudar a língua a partir desta abordagem vem se desenvolvendo desde então na forma de debates e teses (Pennycook, 2004; Sardinha, 2000), além de variadas aplicações, como veremos a seguir.

Leitura vertical

Quando falamos em linguística baseada em corpus, algumas das primeiras pistas que vêm à mente são as linhas de concordância e as listas de palavras geradas por computador na tentativa de extrair sentido aos fenômenos encontrados em coleções de textos. Estas buscas por palavras e expressões em múltiplos contextos datam do século XIII pelas mãos de estudiosos da bíblia que, com o intuito de apontar para seus companheiros as palavras contidas na obra, as indexavam manualmente em arranjos alfabéticos, junto com citações de onde e em que passagens elas ocorriam (McCarthy & O’Keeffe, 2010).

As primeiras concordâncias geradas eletronicamente aparecem em fins dos anos 1950, usando-se tecnologia de cartões perfurados para armazenamento

(Hockey, 2004; McEnery & Hardie, 2012). Avanços consideráveis vieram nos anos 1970, quando cientistas de informação e profissionais de biblioteca desenvolvem interesse pelas concordâncias de palavras-chave em contexto (KWIC - *key word in context*) como uma forma de automatizar análises de assunto, fontes bibliográficas e citações, beneficiados pelos avanços da tecnologia. Porém, são nos anos 1980 e 1990 que os corpora começam a ser tomados para investigações sistemáticas da linguagem: perscrutar textos para encontrar exemplos de um fenômeno particular da língua, melhorar a cobertura de dicionários, avançar em estudos empíricos sobre aspectos da gramática, analisar a linguagem com base em dados reais, etc. (McCarthy & O'Keeffe, 2010).

Enquanto um texto é para ser lido horizontalmente prestando-se atenção às cláusulas, sentenças e parágrafos, um corpus é varrido verticalmente, buscando-se pelos padrões presentes em contexto. O desenvolvimento de ferramentas apoiadas em corpora passa a abranger áreas tão distintas como ensino e aprendizado de línguas, análise do discurso, linguística forense, tradução, pragmática, tecnologia da fala, sociolinguística, entre outras. Nos estudos literários e de tradução, que lidam com comparações de textos e padrões, softwares com recursos para atribuir categorias semânticas às palavras-chave oferecem um imenso escopo para a metodologia de análise estilística.

Na linguística forense, são as características e padrões de tipicidade encontradas e demonstradas estatisticamente, que corroboram com a evidência (ou não) de unicidade ou genuinidade de autoria dos textos. Na pragmática, uma área fértil tem sido o uso de corpora para comparar características como vagueza, ironia, humor, hipérbole e metáfora entre diferentes línguas. Na sociolinguística, os contextos do discurso político e debates parlamentares, assim como coberturas de notícias políticas, levam os linguistas a explorar de forma criativa informações lexicais e morfossintática existentes nos corpora para fazer análise de palavras-chave, análise crítica do discurso e comparações, procurando expor as ideologias subjacentes aos textos (McCarthy & O'Keeffe, 2010).

Assim, embora tenha sua origem vinculada à exploração de fenômenos linguísticos, o trabalho de exploração com corpus vem sendo cada vez mais apropriado por outros campos das humanas, permitindo novas experimentações e

insights, apoiando-se tanto em técnicas quantitativas quanto qualitativas. Mello & Souza (2014) fornecem alguns exemplos de ferramentas e aplicações de corpora voltadas para estudos da língua, mineração de texto e processamento de linguagem natural.

2.3.3 Abordagens

Dentro da linguística com corpus, duas abordagens são comumente associadas aos experimentos de maneira a situá-lo na pesquisa: *corpus-driven* (guiada por corpus) e *corpus-based* (baseada em corpus). Na primeira abordagem o corpus serve como uma base empírica onde dados são extraídos e fenômenos linguísticos identificados sem a existência prévia de expectativas, isto é, conclusões ou afirmações são feitas exclusivamente com base nas observações do corpus. Aqui, há um certo distanciamento do investigador, que parte de uma observação desinteressada para construir suas hipóteses (Freitas, 2015). Linhas de concordância são frequentemente assimiladas nesta abordagem, como lembra Laurence Anthony (2013). Já na *corpus-based*, os pesquisadores possuem de antemão questões iniciais. O corpus é visto como uma espécie de repositório de linguagem e a partir dele dados são extraídos para apoiar um conhecimento prévio intuitivo, verificar expectativas, provar teorias existentes ou permitir que fenômenos linguísticos sejam quantificados ou ilustrados (Freitas, 2015; Archer, 2012).

Seguindo esta mesma direção, Santos (2008) distingue as experimentações com corpus como estudos empíricos de caráter exploratório ou experimental. O primeiro caso colige amostras, conta ocorrências, procura correlações, experimenta classificações e identifica conjuntos. Assim fazendo, abre sendas e identifica campos de interesse, construindo mapas da área que possam ser visitados posteriormente. Já o tipo experimental – em geral produzido com base em explorações anteriores –, possui de antemão uma hipótese que pretende verificar e aferir no corpus (Santos, 2008:49). Como o uso de computador é imprescindível para a realização destas tarefas, Sampson aponta que não é surpresa que grande parte dos linguistas atualmente trabalhe em departamentos da ciência da computação – onde técnicas empíricas são tidas como certas – e não nas de linguística propriamente ditos, tornando evidente a distância criada entre o linguista

de corpus do linguista teórico que não se envolve no exame da linguagem a partir de evidências diretas (2001:6).

Um dos recursos mais comuns de exploração em corpus relaciona-se com o conceito de *bag of words*, que, como a expressão em inglês antevê, olha para todo documento como uma coleção de palavras individuais, ignorando gramática, sintaxe, pontuação ou qualquer outra estrutura linguística. É possível obter, por exemplo, a frequência de palavras para medir quão prevalente um termo encontra-se em um documento, e a frequência inversa do documento para verificar a distribuição desse termo no corpus como um todo (Provost e Fawcett, 2013). Outro recurso interessante são as sequências *n-grams*. *N-grams* são todas as combinações de palavras adjacentes de comprimento *n* que se pode encontrar em um texto, sendo úteis para capturar a estrutura da linguagem do ponto de vista estatístico, tal como que palavra provavelmente seguirá aquela dada.

2.3.4
Tarefa de anotação

Anotar é delimitar um segmento de texto e atribuir-lhe uma etiqueta; é como uma tarefa de classificação. De um certo modo podemos considerar anotações como metadados que proveem informação adicional, ou seja, ao invés de nos dizer o que o texto em si compreende, elas fornecem informação sobre a linguagem desse texto (Leech, 1997).

Corpus anotado contém informações linguísticas que podem ser de natureza variada: informação sintática, gramatical, estilística, semântica, discursivo/pragmática, etc. Estas marcações associadas aos segmentos no texto permitem ao pesquisador ir além das linhas de concordância, facilitando a identificação automática de certas estruturas na forma de listas de distribuição e aumentando consideravelmente as potencialidades de pesquisa.

A construção de um corpus anotado não é uma tarefa trivial de ser cumprida, muito pelo contrário. Os projetos aos quais temos acesso hoje demonstram o significativo esforço intelectual e operacional que a anotação demanda, sendo comum existirem várias versões e revisões do mesmo corpus anotado. Uma das características da anotação é sua dimensão interpretativa, o que pode levar a uma concordância frágil entre diferentes anotadores. Neste sentido, não há nenhuma

maneira puramente objetiva, mecanicista, de decidir a etiqueta que deve ser aplicada a este ou aquele dado fenômeno linguístico (Leech, 1997). Assim, uma coleção dourada que servirá de gabarito não vem para dizer exatamente o que uma palavra ou um sintagma é, mas sim o que se *interpretou sobre* tal palavra ou sintagma.

Da mesma forma que um corpus é sempre compilado segundo certos interesses, a anotação também deve se orientar segundo as demandas e aplicações previstas pela comunidade interpretativa que se beneficiará desse corpus (Hirst, 2009). Interesses linguísticos e interesses de aplicação estão no cerne da questão "para que se quer anotar?", cabendo nessa resposta a definição do tipo de anotação e conjunto de etiquetas a ser adotado.

Leech (1997), ao discorrer sobre estudos com corpus anotado, aponta três razões pelas quais considera relevante o trabalho de anotação: i) extração de informação, que é na verdade a razão de ser de um corpus; 2) reusabilidade, por tornar o recurso disponível para outros usuários; e 3) multifuncionalidade, pelo fato de que as anotações podem ser utilizadas para diferentes propósitos.

No final da década de 1990, o *Penn Treebank* já incluía em seu projeto de anotação, além das classes gramaticais, alguns níveis de classificação sintática, tais como a funcional (identificando, por exemplo, sujeito e objeto), a de tipos adverbiais, de correferência, de papéis semânticos e de sentimentos (Garside et al, 1997). Acrescentar estas informações a um corpus permite ao computador encontrar características capazes de tornar certas tarefas mais precisas (Pustejovsky e Stubbs, 2012). O interesse pode ser, por exemplo, saber a opinião positiva ou negativa que um corpus revelará sobre determinado assunto ou sobre papéis semânticos exercidos pelos sujeitos das sentenças.

O trabalho de anotação pode ser feito de forma manual (em geral, por linguistas), automática (por ferrramentas de PLN) ou semi-automática (com correção manual da saída gerada de outras ferramentas). Segundo Archer (2012), a abordagem automática costuma ser adotada nos estudos linguísticos para anotação gramatical, lematização e anotação semântica, principalmente reconhecimento de entidades nomeadas. A anotação manual costuma ser adotada em corpora de dimensões modestas, e as possibilidades incluem, dentre outras, anotações de

caráter semântico-discursivo-pragmático. Como as categorias linguísticas nesta dimensão são muito dependentes de contexto e sua identificação no texto suscita mais controvérsias que os outros fenômenos linguísticos, a atividade não é trivial para a completa automatização. Não podemos perder de vista que a anotação é, sobretudo, um procedimento que envolve interpretação, classificação e formalização do fenômeno em foco (Santos et al, 2015).

A questão da anotação em corpus não é isenta de debates. Em seu artigo "Corpus annotation: a welcome addition or an interpretation too far?" (2012), Dawn Archer traz a posição de alguns linguistas resistentes à tarefa de anotar um corpus: John Sinclair teme que os usuários passem a observar os dados dos corpus apenas através das etiquetas de anotação, perdendo tudo o mais que elas não conseguem captar; Susan Hunston, por outro lado, aponta que a possibilidade de recuperar dados de forma sistemática, que é um dos pontos fortes de um corpus anotado, pode se tornar uma armadilha se o pesquisador permanecer alheio à possibilidade de que as suas questões de investigação estão sendo modeladas sumariamente pelas categorias utilizadas no processo de recuperação. Esta posição, segundo Archer, pode ser inclusive distorcida quando tomada por aqueles contrários à abordagem *corpus-based* ao sugerirem que certos pesquisadores, tendo uma teoria já pré-concebida, podem deliberadamente ignorar ou descartar dados que vão contra essa teoria. Ou seja, podem escolher trabalhar apenas com aquilo que se "encaixa" no projeto de investigação (Archer, 2012). No entanto, para Archer este argumento não se sustenta se o esquema de anotação for pensado e concebido a priori como um meio através do qual queremos examinar as evidências, e só então ao examinar essas evidências, determinar se elas se encaixam ou não em alguma noção pré-concebida.

Para Archer, sua experiência com anotação em corpus sempre a permitiu "ver as coisas em novas perspectivas", não importa quais sejam. Para ela trata-se de um acréscimo muito bem-vindo à pesquisa, desde que não percamos de vista como os textos são moldados por seus autores e, sobretudo, o período em que foram produzidos (Archer, 2012).

Cláudia Freitas enxerga aí uma oportunidade para que a leitura distante deixe de ser vista apenas como uma abordagem meramente quantitativa. Ainda que

até o momento não seja uma estratégia comum, a anotação pode adicionar mais uma camada a este tipo de leitura, trazendo uma dimensão qualitativa ao trabalho, na medida em que se trata de uma atividade classificatória por excelência que procura atribuir sentidos mais gerais às ocorrências únicas no texto (Santos et al, 2020:290).

Sinclair aponta que o valor de um corpus não é medido propriamente pelo seu tamanho, mas por critérios que podem ir da diversidade, representatividade e balanceamento – isto é, equilíbrio de gêneros discursivos, tipos de texto etc. – segundo algum propósito, até o grau de informação adicional agregado na forma de anotações (Sinclair, 2005). O valor da anotação pode ser mais bem compreendido se a considerarmos tão somente como uma forma de marcar características no texto que não são imediatamente observáveis a olho nu (Anthony, 2013:148). No final, a importância está no tipo de informação que dele podemos extrair com o auxílio da ferramenta certa (Anthony, 2013:149).

2.4
Extração de informações

Quando falamos de extração de informações (IE) nos referimos ao processo de obtenção automática de estruturas – tais como entidades, relações entre entidades e atributos que descrevem entidades – a partir de fontes não estruturadas. A tarefa busca formas mais ricas para recuperar informação, do que seria possível apenas utilizando pesquisas por palavras-chave. A partir de técnicas que descreveremos mais adiante, um sistema IE pode, por exemplo, identificar e extrair de uma coleção de textos o nome de todas as organizações mencionadas, incluindo aquelas que o usuário não detinha conhecimento prévio, e ainda, o nome de todas as pessoas que possuem algum vínculo com essas organizações e o tipo de vínculo.

Há mais de três décadas a área tem mobilizado uma verdadeira comunidade de pesquisadores de ramos como representação do conhecimento, linguística computacional e inteligência artificial, que se lançam a esse desafio de conceber métodos automáticos para identificar dados relevantes em meio a textos livres (Sarawagi, 2007; Grishman, 2015).

2.4.1
Evolução do campo

A IE tem sua escalada na comunidade de processamento de linguagem natural com o ímpeto dos fóruns de competição iniciados nos anos 80 e 90s. Nestes eventos, os organizadores definem um desafio específico de IE/NLP e métricas para a avaliação de desempenho dos sistemas concorrentes. Em geral, disponibilizam uma coleção dourada para servir de gabarito, além de materiais de treinamento, e no final comparam a efetividade das abordagens empregadas pelos participantes.

Alguns se mantêm ativos até hoje, mas muitos foram descontinuados ou substituídos por versões mais atuais. Dentre os mais importantes temos: o MUC (*Message Understanding Conference*), o ICDAR (*International Conference on Document Analysis and Recognition*), o CoNLL (*Conference on Computation Natural Language Learning*), o ACE (*Automatic Content Extraction*), o HAREM (Avaliação de Reconhecimento de Entidades Mencionadas), a TAC (*Text Analysis Conference*), e mais recentemente, a IberLEF (*Iberian Language Evaluation Forum*).

Estes fóruns resultam em material valioso para estudos da área e nos dão acesso aos modelos semânticos e as diretrizes fornecidas pelos organizadores, além das metodologias, técnicas e recursos produzidos pelos concorrentes.

O MUC[5] é o mais antigo, tendo a primeira conferência ocorrida em 1987, ocasião em que a tarefa de extração de informação ainda não possuía diretrizes bem definidas. A partir da sexta edição (MUC-6), em 1995, o reconhecimento de entidades nomeadas (NER) passou a fazer parte da conferência (Grishman & Sundheim, 1996). O CoNLL[6] tem suas origens em 1997 com foco em *machine learning* para o processamento de linguagem natural, trazendo nas edições de 2002 e 2003 a tarefa de NER para o centro da avaliação (Marrero et al, 2009). O ACE[7] foi gestado pelo NIST (National Institute of Standards and Technology, EUA) entre os anos de 1999 a 2008, como um programa de pesquisa para o desenvolvimento de tecnologias de extração de informação. Segundo Maynard e colegas, teve a

[5] http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html e
http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html
[6] http://www.conll.org/previous-editions
[7] https://www.ldc.upenn.edu/collaborations/past-projects/ace

intenção de melhorar as diretrizes do MUC, valorizando a análise semântica em detrimento da análise puramente formal (Maynard et al, 2003, apud Santos, 2007). Também priorizou o lado do *conteúdo* e propôs que a identificação de entidades pudesse ser realizada sem restrições de forma, isto é, não apenas como nomes próprios, mas também como substantivos comuns, pronomes e sintagmas nominais (Santos, 2007). Além disso, o ACE buscou combinar a tarefa de reconhecimento de entidades com a de co-referência (Santos, 2007), e incorporou expressões temporais e anotação de eventos como tarefas independentes (Marrero et al, 2009).

O fórum de avaliação HAREM, organizado pela Linguateca[8], é específico para a língua portuguesa e teve duas edições: a primeira entre 2004 e 2006 (Santos & Cardoso, 2007), e a segunda entre 2007 e 2008 (Santos, 2008a). Segundo os organizadores, o modelo semântico do HAREM assenta em dois aspectos essenciais, que o distingue de outros modelos utilizados na avaliação de NER: i) a identificação e classificação de uma entidade depende do seu uso em contexto e; ii) é possível atribuir mais de uma classificação a uma mesma entidade.

O TAC[9] é uma série organizada desde 2008 pelo NIST para incentivar a pesquisa em PLN e aplicações relacionadas, fornecendo uma grande coleção de testes, procedimentos de avaliação e um fórum para os participantes compartilharem seus resultados (NIST, 2019). O TAC compreende conjuntos de tarefas conhecidas como "trilhas", cada uma enfocando um subproblema específico de PNL que são definidas a cada edição. As de 2020, por exemplo, envolveram sistemas de perguntas e respostas sobre a COVID-19, reconhecimento de nomes, menções nominais e pronominais de entidades em artigos de notícias e extração de conhecimento a partir de fontes multimída.

Por fim, o mais recente de todos, o IberLef[10], estreou em 2019 com o objetivo de organizar tarefas competitivas para a comunidade de processamento de linguagem natural para as línguas ibéricas espanhol, português, catalão, basco e galego. Além das usuais tarefas de reconhecimento de entidades nomeadas e extração de relações, também foram incluídas descoberta de conhecimento em áreas

[8] https://www.linguateca.pt/
[9] https://tac.nist.gov/
[10] https://sites.google.com/view/iberlef-2019

específicas, identificação de humor e ironia, detecção de autoria, análise de sentimentos e outras tarefas de classificação automática de texto.

Além dos fóruns, um mapeamento das iniciativas neste campo mostra que desde a década de 1980 e principalmente na década de 1990 vem-se produzindo trabalhos relevantes sobre aquisição automática de informação a partir de corpora (Coates-Stephen, 1992; Kupiec, 1993; Hearst, 1992, 1998). Já então se investigavam estratégias para a identificação de relações lexicais a partir de dicionários online (Nakamura, 1988) e textos livres compilados (Jacobs & Zernik, 1988; Hearst, 1992, 1998); se buscavam soluções para lidar com o problema da ambiguidade dos nomes próprios (Coates-Stephen, 1992); se construíam protótipos de sistemas de perguntas e respostas (Kupiec, 1993), dentre muitas outras aplicações. Os componentes responsáveis pela análise linguística de então são fortemente apoiados no exame dos contextos léxico-sintáticos circunscritos aos sintagmas-alvo, ou seja, nos padrões textuais em que determinada palavra, estrutura ou expressão específica pode ocorrer (Hearst, 1992, 1998).

2.4.2
Métodos de PLN

Quando tratamos de dados não estruturados no contexto desta tese estamos nos referindo basicamente a textos. Neste caso, muitos dos métodos da linguística computacional, ou processamento de linguagem natural, são aplicados no processo de extração de informação no corpus. Nesta seção enumeramos as mais comuns (Manning & Schutze, 1999; Santos, 2007; Jurafsky & Martin, 2009)

i) Tokenização. Na análise léxica, a tokenização (ou atomização) é o processo de quebrar o fluxo de texto em unidades de informação. Em geral, figura como um dos momentos mais básicos, anterior às demais etapas. Um token pode ser uma palavra, número, data, url, símbolo ou outro elemento significativo, e normalmente sua delimitação é identificada através de espaços em branco, caracteres de controle ou tokens delimitadores. Um dos cuidados nesta tarefa é identificar corretamente as fronteiras de um token, como nos casos de clíticos, hifens, pontos em abreviação de palavra etc.

ii) Segmentação de sentenças. O processo de segmentar um texto em sentenças em geral se faz baseado no ponto final, ponto de exclamação ou

ponto de interrogação. Os dois últimos são marcadores relativamente inequívocos, mas o ponto final precisa ser olhado com atenção de forma a não o confundir como parte de um token, no caso de abreviações. Por isso, a segmentação de palavras e sentenças são tarefas que tendem a ser endereçadas em conjunto, e cada sistema parseador tem suas próprias regras ou classificadores.

iii) Análise léxica: *stemming* e lematização. Esta etapa opera a nível da palavra e lida com o processamento morfológico e redução de variantes. No primeiro caso, o objetivo é extrair o radical das palavras, suprimindo as flexões que indicariam formas verbais ou de número. Assim, {sonh} é o radical comum de {sonhou, sonhador, sonho, sonhos}, obtido por um processo heurístico que suprime o final das palavras. Já o segundo caso é um tipo de normalização que reduz as variações de uma determinada palavra à sua forma lexical de base, em geral a que encontramos no dicionário. Por exemplo, {correm, corri, correram} são flexões do lema {correr} e a redução inclui todas em uma única busca no corpus.

iv) Marcação gramatical (*POS tagging*). Trata-se da etapa de identificação das classes das palavras, também conhecidas como classes morfológicas, etiquetas lexicais ou partes de falas ('Part-Of-Speech'). De acordo com o contexto em que cada palavra comparece na sentença, uma informação linguística de classe gramatical lhe é atribuída, podendo ser, por exemplo, verbo, substantivo, adjetivo, pronome etc. Na análise sintática a estrutura gramatical de uma frase é importante para determinar as relações entre palavras. Os conjuntos de marcações gramaticais variam conforme o parser e o idioma, e um dos principais desafios reside no tratamento de palavras ambíguas no processo de análise, tal como a atribuição de 'paciente' como adjetivo ou como substantivo.

Análise sintática

As abordagens adotadas para a análise sintática em PLN variam dentre aquelas que se valem de regras e as que utilizam estatística, havendo sistemas que se apoiam em ambas. A primeira é baseada principalmente em gramáticas livres de contexto (*context-free gramars*), por exemplo, o modelo bastante conhecido HPSG

head-driven phrase-structure grammars que tenta fazer a correspondência entre uma gramática e uma determinada frase. Abordagens estatísticas também usam gramáticas, porém tentam induzir gramáticas com base em modelos probabilísticos, por exemplo, PCFG *probabilistic context-free grammars* (Bick, 2000; Jurafsky & Martin, 2009).

A saída gerada pela análise sintática pode ser uma estrutura hierárquica da frase de entrada – chamada de árvore de sintaxe ou estrutura constituinte –, na sua decomposição em sintagmas (nominais, verbais, preposicionais etc.) ou uma estrutura de dependência que estabelece relações binárias entre as palavras – mesmo distantes –, que por sua vez ganham etiquetas de acordo com o papel que exercem, podendo ser sujeito, objeto etc.

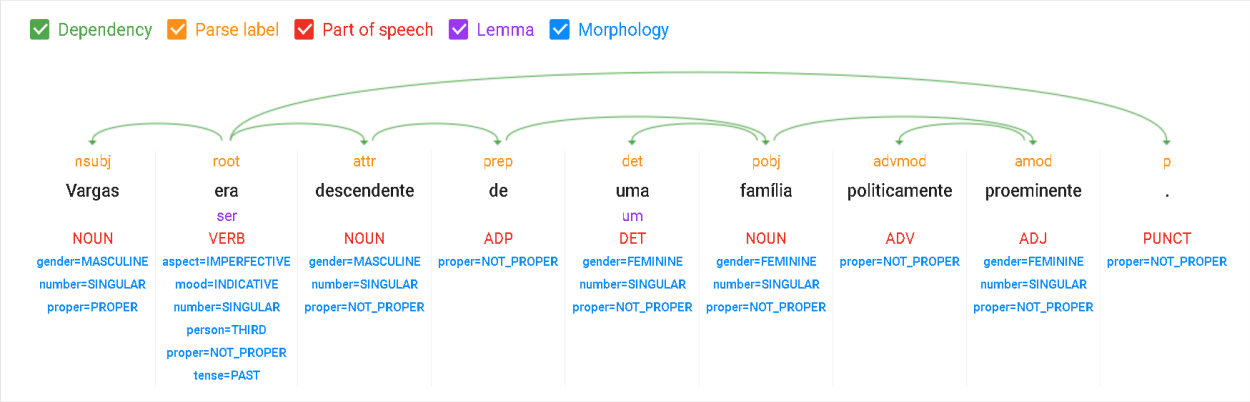

Figura 3 - Estrutura de dependência de uma sentença (obtida com o Google Cloud Natural Language Services)

A figura 3 mostra as marcações gramaticais e as relações sintáticas presentes em uma dada sentença extraída do verbete Getulio Vargas, do DHBB[11].

2.4.3 Tarefas de IE

Esta seção apresenta de forma resumida algumas das principais subtarefas relacionadas à extração de informação.

[11] É importante destacar que as análises não são isentas de discussões. A sentença de exemplo pode ser encontrada aqui: http://www.fgv.br/cpdoc/acervo/dicionarios/verbete-biografico/getulio-dornelles-vargas

Reconhecimento de entidades nomeadas (NER)

O termo ‘entidade nomeada’ (ou mencionada) é a adaptação do conceito inglês “named entity” (NE) cunhado originalmente durante a sexta edição do MUC. Percebeu-se, à época, que para perguntas sobre ‘quem, onde e quando’ era necessário identificar certas unidades de informação presentes nos corpora – como nomes próprios do tipo pessoa, organização, lugar e tempo –, levando ao estabelecimento desta que se tornaria uma importante subtarefa de IE (Grishman & Sundheim, 1996; Santos e Cardoso, 2007; Marrero et al, 2012). Desde então os tipos de entidade se expandiram para atender às mais diversas aplicações e domínios do conhecimento, e hoje encontramos modelos de classificação que incluem toda sorte de tipos, como nomes de eventos, doenças, monumentos, obras, veículos de comunicação etc.

A noção de entidade deve ser um conceito estável e razoavelmente partilhado. Segundo Santos (2007):

> *“...delimitar o conceito de entidade mencionada, como conceito semântico, tem a ver com a relação entre a língua e o mundo exterior à língua, mundo esse que é mediado/representado por um conjunto de símbolos que representam esse mundo. A tarefa de REM*[12]*, como qualquer tarefa semântica, passa por um conjunto de categorias, sobre as quais se tenta chegar a um entendimento partilhado.”* (Santos, 2007:44)

Cada sistema define como as entidades deverão ser tratadas e identificadas no processo de NER, não existindo certo ou errado nesta definição. Naturalmente, quando já há uma compreensão do que esperar da extração, essas decisões terão impacto nos resultados. A seguir, reunimos alguns exemplos do que seriam estes entendimentos partilhados, retirados principalmente dos fóruns de avaliação já mencionados na seção 2.4.1.

No MUC, o tipo ENAMEX foi criado para identificar entidades concernentes às organizações, pessoas e locais, deixando claro que o modelo limita a seleção aos nomes próprios e acrônimos, independente do contexto em uso: por

[12] Reconhecimento de entidade mencionada, equivalente ao NER

exemplo, no caso de “the Clinton government”, apenas *Clinton* será selecionado e receberá a atribuição *person* (pessoa); em “U.S. exporters”, a palavra *exporters* será descartada e *U.S.* identificada como *location* (local). No caso de expressão com mais de uma palavra ou coordenada, a diretriz prega a identificação de uma única entidade, como em: <ENAMEX TYPE = “LOCATION”>*North and South America*</ENAMEX> (Chinchor, 1997).

Já no ACE, os anotadores são orientados a capturar todas as menções de cada entidade no texto, sejam elas indicadas por um nome próprio, por um sintagma ou um pronome. Assim, são aceitos como entidades do tipo *person*: “Joe Smith”, “the guy wearing a blue shirt”, “he, him” (LDC, 2008:4). E ao contrário do MUC, no ACE não se adota a classificação estrita, entendendo que em casos como no trecho “The US navy now says that...”, toda a expressão *The US navy* deve ser considerada uma entidade, do tipo ORG-GOV (organização-governo) (LDC, 2008:49).

O HAREM determina em suas diretrizes a existência do critério formal de maiúscula na identificação das entidades (“médio oriente”, por exemplo, não deve ser considerado), mas ressalta: expressões onde minúsculas claramente fazem parte de uma NE devem ser incluídas, como em “ministro da Administração Interna” (pessoa/cargo) ou “relógio de Sol” (coisa/classe). É preciso ter cuidado nos casos de sintagmas nominais, como “a casa do João”, onde apenas João deve ser marcado (Santos et al, 2008:279). Uma estratégia adotada pelo HAREM foi elaborar e fornecer aos competidores uma lista de palavras ou expressões em minúsculas que devem ser consideradas parte de uma NE para os tipos pessoa/individual, pessoa/grupoind, abstraccao/estado e coisa/substancia. Assim, por exemplo, “conde”, “duque”, “governador”, “presidente”, “rei”, devem ser capturados junto ao nome da pessoa; “doença”, “síndrome”, “mal”, etc, devem se associar ao nome de doenças, e assim por diante (Santos et al, 2008:285).

Entidades encaixadas

Com frequência encontramos entidades que parecem encaixadas umas nas outras, tal como em “Colégio de Aplicação da Universidade Federal de Pernambuco”. Seria uma única entidade? Duas (“Colégio de Aplicação” e

'Universidade Federal de Pernambuco")? Ou três, se considerarmos o estado separadamente ("Pernambuco")?

Nas diretrizes do MUC, a orientação para casos assim está em "Entity-expressions that possess valid entity-expressions", que diz que nas construções que remetem à noção de pertencimento, as sequências envolvidas devem ser marcadas separadamente. Dessa forma, no exemplo "Temple University's Graduate School of Business" marcamos duas entidades do tipo organização: *Temple University* e *Graduate School of Business*. Em "Canada's Parliament", temos *Canada* como local e *Parliament* separado como uma organização (Chinchor, 1997).

No ConLL-2003 diz-se que as entidades não se sobrepõem, ou seja, no caso de uma entidade estar possivelmente encaixada em outra, apenas a entidade de nível mais superior será marcada (Sang, 2002; Sang e De Meulder, 2003), como em "IX Asamblea General de la Alianza Mediterránea de Agencias de Noticias", "Reserva Federal de Estados Unidos" ou "XIX Feria de Muestras de la Campiña Sur de Extremadura".

No Primeiro HAREM, a orientação aos participantes apontava no sentido de marcarem preferencialmente a NE mais longa (Santos e Cardoso, 2007). Dessa forma, a expressão "presidente da Câmara de Nova York" seria considerada uma única entidade da classe pessoa, tipo cargo. No Segundo HAREM, no entanto, os organizadores decidiram aplicar a etiqueta "alt" como um recurso para representar alternativas nas estruturas passíveis de constituir mais de uma NE. Assim, a mesma expressão ganha anotações diversas:

- "presidente da Câmara de Nova York" – pessoa, tipo cargo
- "Câmara de Nova York" – organização, tipo administração
- "Câmara" – organização, tipo administração
- "Nova York" – local, tipo humano, subtipo divisão

Segundo Carvalho et al (2008), este procedimento pode ter interesse a vários níveis. Além de permitir uma análise mais fina sobre o próprio mecanismo de composição de certas entidades, possibilita a identificação de NER que, de outro modo, não seriam analisadas (Carvalho et al, 2008:19).

Nas diretrizes do ACE casos parecidos são tratados como *nested premodifiers*, ou seja, correlacionando os elementos capazes de modificar o

valor/significado de outros elementos na estrutura. Por exemplo, em "White House press secretary Scott McClellan", existe a presença tanto de pré-modificadores encaixados quanto não encaixados. *White House* e *press* modificam *secretary* mas não se encaixam um com ou outro. Dessa forma, a expressão permite a identificação de quatro entidades:

- "White House" – organização, tipo governamental
- "Press" – organização, tipo comercial
- "White House press secretary" – pessoa
- "Nova York White House press secretary Scott McClellan" – pessoa

Apresentamos abaixo um quadro comparativo sintético com as decisões relatadas acima:

Fórum de avaliação	Definição de entidade	Aceita entidades encaixadas?	Classes principais	Aceita classificação múltipla?
MUC-7	Nomes próprios	Não	7 (pessoas, organizações, locais, data, tempo, moeda, porcentagem)	Não
ConLL03	Nomes próprios	Não	4 (pessoas, organizações, locais, miscelânea)	Não
ACE	Nomes próprios, substantivos comuns, pronomes, sintagmas nominais etc	Sim, considerando os modificadores	7 (pessoas, organizações, locais, entidades geopolíticas, instalações, veículos e armas)	Não (mas aceita múltiplas vertentes)
HAREM 2	Nomes próprios e substantivos comuns	Sim	10 (pessoas, organizações, locais, obras, valor, tempo, acontecimentos, coisas, abstrações e outros)	Sim

Tabela 1 – Quadro comparativo dos principais fóruns de avaliação de REM

Lembrando que as definições exemplificadas no quadro foram estabelecidas para guiar a avaliação dos sistemas participantes dos fóruns apontados. Cada sistema de NER deve planejar seu modelo de categorias a partir dos objetivos do projeto em questão.

Léxicos

Léxicos são atalhos rápidos para a identificação de entidades nomeadas no corpus, ou seja, ao encontrar uma palavra que pertence a uma determinada lista, o sistema anota com a etiqueta da classe atribuída. Uma das motivações para sua criação é o fato de que os parsers não são capazes de identificar com 100% de precisão certos termos, mesmo com a ajuda de heurísticas morfossintáticas aplicadas no reconhecimento. Por exemplo, na saída abaixo produzida pelo analisador PALAVRAS da sentença "a Suemi apresentará um seminário na semana que vem.":

```
a [o] <artd> DET F S @>N #1->2
Suemi [Suemi] <org> <heur> PROP F S @SUBJ> §AG #2->3
apresentará [apresentar] <fmc> <vH> <mv> V FUT 3S IND VFIN @FS-STA §PRED #3->0
um [um] <arti> DET M S @>N #4->5
seminário [seminário] <inst> <occ> N M S @<ACC §PAT #5->3
em [em] <sam-> PRP @<ADVL #6->3
a [o] <-sam> <artd> DET F S @>N #7->8
semana [semana] <dur> N F S @P< §LOC-TMP #8->6
que [que] <clb> <clb-fs> <rel> SPEC M S @SUBJ> §TH #9->10
vem [vir] <mv> <np-close> V PR 3S IND VFIN @FS-N< §ATR #10->8
. PU @PU #11->0
```

Tabela 2 - Reconhecimento de entidades nomeadas pelo PALAVRAS (obtido no VISL)

A pessoa "Suemi" foi erroneamente identificada como organização <org> e "seminário" ganhou duas categorizações: uma como instituição <inst> e outra, aparentemente mais correta, como evento organizado <occ>.

Apesar da praticidade, uma das desvantagens no uso de listas é que elas não lidam com variantes de nomes e nem resolvem ambiguidade: "Getúlio Vargas" e "Getúlio Dornelles Vargas" seriam consideradas duas pessoas diferentes; já "Washington" poderia ser lugar, pessoa ou até uma organização, mas só saberíamos com certeza em uma leitura atenta do trecho onde o termo se localiza. Segundo Mikheev e colegas (1999), a adoção de listas nem sempre corresponderá a um grande ganho porque nomes de pessoas, organizações e outras NEs fazem parte de uma classe de palavras conhecida como *open word class*, isto é, há sempre novas instâncias a acrescentar.

Extração de relações semânticas

Descobrir quais indivíduos nasceram em quais lugares, frequentaram quais instituições e exerceram quais cargos, necessariamente conduz à identificação das relações semânticas entre as entidades nomeadas.

A tarefa de identificar estas relações constitui-se um desafio na área de IE, tendo em vista o conhecimento linguístico e a sofisticação das técnicas de processamento da língua exigidos. Em geral, o tipo de relações extraídas é altamente dependente do domínio, pois corresponde ao que se deseja obter de informações, por exemplo, cidades (*localizadas-em*), pessoas (*casado-com*, *nascido-em*) e organizações (*subsidiária-de*, *funcionário-de*)[13].

Relações podem ser representadas no formato de triplas, como em < *e1*, *rel*, *e2*>, onde *e1* e *e2* são os sintagmas nominais/entidades de uma relação e *rel* é o tipo de relação que está conectando os dois sintagmas. Algumas das relações semânticas mais comuns estão no domínio lexical: a hiponímia/hiperonímia (relação *é-um*), meronímia/holonímia (relação *parte-de*), sinonímia e causa-efeito. As relações de hiponímia e meronímia, em especial, são bastante usadas para a expansão de consultas e na construção ou melhoria de ontologias e bases léxicas (Hearst, 1992, 1998; Freitas, 2007). Por exemplo, na sentença "alguns dos Atos Institucionais, como o AI-2 e o AI-5, cercearam os direitos...", duas relações de hiponímia podem ser extraídas: <AI-2, *é-um*, Ato Institucional> e <AI-5, *é-um*, Ato Institucional>. Da mesma forma, na sentença "O Congresso Nacional, constituído de suas duas casas, o Senado Federal e a Câmara dos Deputados...", podemos encontrar as seguintes relações de meronímia: <Senado Federal, *parte-de*, Congresso Nacional> e <Câmara dos Deputados, *parte-de*, Congresso Nacional>.

Para ilustrar, na sentença "Getúlio Dornelles Vargas, filho de Manuel do Nascimento Vargas, nasceu em São Borja", primeiro identificamos as entidades mencionadas, no caso as três em destaque:

- Getúlio Dornelles Vargas = Pessoa
- Manuel do Nascimento Vargas = Pessoa
- São Borja = Lugar

[13] Um conjunto comum de relações é fornecido pelo Programa Linguistic Data Consortium in the Automatic Content Extraction: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-relations-guidelines-v6.2.pdf

Em seguida, as duas relações presentes, *filho-de* e *nasceu-em*:

- <Getúlio Dornelles Vargas, *filho-de*, Manuel do Nascimento Vargas>
- <Getúlio Dornelles Vargas, *nasceu-em*, São Borja>

A extração de relações requer processamento prévio do texto, o que envolve análise sintática e NER. Os principais métodos podem ser discriminados em abordagens semelhantes às usadas no domínio de reconhecimento de entidade nomeada, baseadas em padrões codificados manualmente ou métodos de aprendizado de máquina. Na seção 2.4.4 apresentaremos as abordagens mais comuns.

Resolução de correferência

Podemos definir correferência a relação entre termos que se referem a uma mesma entidade de mundo e a sua resolução busca evidenciar esta relação (Grishman, 2015).

Mas para compreender o fenômeno, é preciso também trazer para o palco o conceito de anáfora, já que ambos estão relacionados. Sem entrar em discussões linguísticas mais aprofundadas, podemos definir anáfora como a retomada de um referente que foi apresentado anteriormente no texto, funcionando como uma espécie de substituto daquele (Marcuschi, 2001). Quando uma entidade é mencionada pela primeira vez, temos a *evocação* da entidade, e quando ela é mencionada mais tarde, temos a realização do *acesso* a essa entidade. A expressão que faz o acesso é chamada anafórica e a expressão anterior é o seu antecedente. A relação entre essas duas expressões (anáfora e antecedente) é chamada de relação de correferência (Vieira et al, 2008; Jurafsky & Martin, 2009).

As formas mais comuns de anáfora no contexto de resolução em PLN são aquelas realizadas por pronomes anafóricos (1) e (2); por sintagmas que contenham determinantes (3); ou quando ambas as expressões de anáfora e antecedente possuem o mesmo nome-núcleo (4) ou são quase sinônimos (5):

(1) Francisco Campos viajou então a Porto Alegre para verificar in loco os preparativos da revolução. Com Vargas e Osvaldo Aranha, ele acertou o esquema de participação de Minas no levante.

(2) Francisco Campos viajou então a Porto Alegre para verificar in loco os preparativos da revolução. Seus companheiros acertaram o esquema de participação de Minas no levante.

(3) O político não obteve sucesso na eleição para a Câmara dos Deputados daquele ano. Mas disse que tentaria novamente quando fosse realizada uma nova.

(4) Em 1º de junho, Vargas lançou seu Manifesto à Nação. Nesse manifesto, ele ataca as bancadas mineira e paraibana...

(5) Em 1º de junho, Vargas lançou seu Manifesto à Nação. O texto ataca as bancadas mineira e paraibana...

Concordâncias numérica (singular, plural) e de gênero (feminino, masculino) entre anáfora e antecedente, além da consistência semântica das orações em que aparecem, são algumas das pistas que os sistemas de resolução de correferência utilizam (Cambria et al, 2015).

Os atuais sistemas de resolução utilizam recursos de inteligência artificial, como o aprendizado de máquina, para atacar o problema. Segundo Vieira e colegas, em muitas dessas abordagens um conjunto de características linguísticas, identificadas num corpus previamente anotado com correferência, é analisado automaticamente para que se identifiquem relações e presença de ligação entre as expressões (Vieira et al, 2008).

Alguns modelos usam a informação estatística representada pelas frequências dos padrões obtidos de um corpus selecionado e a aplicação de heurísticas para encontrar o candidato antecedente com a frequência mais alta (Cambria et al, 2015).

Há muitos trabalhos dedicados à resolução de correferência na comunidade de PLN. Mas a tarefa não é trivial, pelo contrário, ainda é um desafio grande para a computação linguística, a despeito dos muitos avanços percorridos na área. O problema impacta diretamente na construção da coerência de um texto e, por consequência, na qualidade dos resultados da extração automática de informações.

2.4.4
Abordagens para IE

Segundo a literatura (Curran & Clark, 2003; Smith & Osborne, 2006; Sarmento et al, 2006; Romão, 2007), muitos dos sistemas de NER funcionam sobre duas estratégias: utilizando regras de inferência a partir de pistas linguísticas e utilizando listas lexicais. As pistas viabilizam a identificação de padrões e podem se basear em características diversas, tais como, morfossintáticas, ortográficas, de contexto etc. e serem dependentes ou não do idioma. Já as listas lexicais, como vimos anteriormente, são mais simples e não consideram os contextos em que os termos aparecem, ou seja, não tratam ambiguidades.

Os métodos para detectar relações semânticas e extrair informações de textos podem ser grosseiramente classificados em: i) abordagens baseadas em regras; ii) abordagens de aprendizado de máquina (ML) supervisionadas, semi supervisionadas e não supervisionadas; e iii) abordagens híbridas.

Abordagens baseadas em regras

Na primeira abordagem, regras de extração são desenvolvidas manualmente e dependem principalmente de aspectos léxico-sintáticos de sentenças. Ainda que à primeira vista sejam construções que se apresentam de forma muito simples na língua, a formalização destes padrões traz resultados positivos ao servirem como pistas para a descoberta automática de estruturas de informação no texto.

Segundo Hearst (1992), diferentes relações podem ser expressas utilizando um pequeno número de padrões léxico-sintáticos. É dela a experiência mais antiga e conhecida para extrair relacionamentos do tipo *is-a* (hiperônimos). Esses padrões são construídos a partir de frases contendo pistas como "*and*", "*such as*", "*like*", "*or*" etc., combinadas à sinais de pontuação, marcadores de posição para entidades nomeadas e elementos de expressões regulares. Por exemplo, o padrão "such NP as {NP ,}* {(or | and)} NP" pode ser aplicado em exemplos do tipo "*...works by such authors as Herrick, Goldsmith, and Shakespeare*", extraindo as seguintes relações: "hyponym("author", "Herrick"), hyponym("author", "Goldsmith"), hyponym("author", "Shakespeare") (Hearst, 1992, 1998).

A principal vantagem do desenvolvimento de regras é que, devido à sua natureza declarativa, esses padrões são compreensíveis pelos humanos e os efeitos da mudança são diretamente visíveis se comparado com um modelo de aprendizado de máquina, que requer uma fase de treinamento e uma fase de extração. Em geral, a qualidade da informação extraída é bastante alta, mas em contrapartida os níveis de abrangência costumam ser baixos (Makarov, 2018:104).

Essa abordagem tem como principais questões a escalabilidade, diante dos altos custos de desenvolvimento de regras, e o gerenciamento de grandes conjuntos de regras (Makarov, 2018).

Abordagens de aprendizado de máquina

O método supervisionado de aprendizado de máquina é baseado em características lexicais, sintáticas e semânticas extraídas a partir de dados de exemplos previamente etiquetados (Jurafsky & Martin, 2009). Estes dados funcionam como gabarito (golden) e são usados para treinar algoritmos de classificação que atribuem o tipo de relacionamento mais provável às respectivas menções de entidade. Essas características podem ser baseadas em palavras (por exemplo, palavra antes e depois das entidades), gramaticais (elemento raiz de frases) ou semânticas (tipos de entidade, classes gramaticais de palavras dependentes das entidades etc.). A vantagem desta abordagem é que a validação cruzada pode ser facilmente aplicada para avaliar as características. A desvantagem é que é necessária uma amostra razoavelmente grande de exemplos anotados, o que pode ser custoso (Sarawagi, 2008).

Na extração semi supervisionada a ideia é reduzir o alto esforço de criação de dados anotados para treinamento. A ideia é usar algoritmos de bootstrapping, onde uma pequena quantidade de exemplos de instâncias de relação, chamadas sementes, são fornecidas como pontapé inicial (Golshan et al, 2018). Por exemplo, se estivermos procurando por autores de livros, usamos a tupla {"JK Rowling", "Harry Potter e a Pedra Filosofal"}, {"Stephen King", "Misery"}. As ocorrências desses exemplos são pesquisadas em um grande corpus cru e os padrões são aprendidos olhando-se para essas ocorrências. Padrões recém-descobertos são usados para extrair novas instâncias de relação e, em seguida, o processo é repetido. O principal desafio deste processo é a avaliação dos padrões descobertos, pois não

existe um padrão que funcione como gabarito. Havendo divergência semântica quanto à um padrão, a introdução de tuplas errôneas podem levar à sucessão de padrões erráticos no processo de iteração. Para tentar resolver esse problema, valores de confiança são estimados para novos padrões e novas tuplas com base em quantas tuplas o padrão encontra no conjunto já extraído e em toda a coleção de documentos (Agt-Rickauer, 2019). Recursos linguísticos como tesauros e wordnets também podem ser usados para melhorar o desempenho da tarefa (Taba, 2013).

Por fim, o método de aprendizado de máquina não supervisionado é a abordagem que visa extrair relações e instâncias de relações sem qualquer intervenção humana. A maior diferença, se comparado aos métodos anteriores, é que ele não se restringe a um conjunto fixo de relações. Na maioria dos casos, a abordagem recai no processo de clusterização (agrupamento) para determinar conjuntos de pares de entidades que pertençam ao mesmo tipo de relação. Em geral, o procedimento passa pelas seguintes etapas: i) reconhecimento de entidade nomeada no corpus inteiro; ii) pares de entidades co-ocorrentes, sua ordem e seu contexto (palavras intermediárias e vizinhas) são registrados; iii) semelhanças de contexto são determinadas (por exemplo, por frequências de palavras, informações lexicais e estruturas de dependência); iv) o agrupamento hierárquico é aplicado aos pares de entidades usando valores de similaridade (Agt-Rickauer, 2019). No final, cada um dos agrupamentos resultantes representa uma relação, que é então rotulada.

Um importante método usado para construir automaticamente representações semânticas de palavras e frases observando sua ocorrência no corpus é conhecido como Semântica Distribuída (Distributional Semantics). Esse método assume que a distribuição estatística das palavras em contexto desempenha um papel fundamental na caracterização de seu comportamento semântico. Os modelos distributivos têm uma longa história em pesquisa linguística, cognitiva e computacional que remonta aos anos 1950. Em seus primeiros trabalhos sobre análise do discurso, Zellig Harris cunhou pela primeira vez em 1954 a hipótese distributiva: “as partes de uma linguagem não ocorrem arbitrariamente em relação umas às outras: cada elemento ocorre em certas posições em relação a certos outros elementos” (Harris, 1954). O famoso slogan de J.R. Firth: “You shall know a word

by the company it keeps"[14] mostra exatamente isso: palavras com significados semelhantes ocorrem em contextos semelhantes.

Vector Space Models ou VSM (Modelos de Espaço Vetorial) são a representação mais comum de contextos de palavras. Contextos, nesse caso, são as palavras adjacentes a uma palavra-alvo. A ideia geral do VSM é representar documentos, frases ou palavras como pontos no espaço, e a distância entre esses pontos corresponde à sua semelhança semântica (Taba, 2013; Golshan et al, 2019). Essa medida geométrica foi generalizada para capturar vários aspectos de co-ocorrência em textos de linguagem natural, geralmente com o uso de vetores, matrizes e tensores.

Os modelos de espaço vetorial podem ser resumidos como matrizes de co-ocorrência de palavras ponderadas com informações mútuas.

	bite	buy	drive	eat	get	live	park	ride	tell
bike	0	9	0	0	12	0	8	6	0
car	0	13	8	0	15	0	5	0	0
dog	0	0	0	9	10	7	0	0	1
lion	6	0	0	1	8	3	0	0	0

Figura 3 - Matriz de co-ocorrência de palavras (reproduzido de Agt-Rickauer, 2019:50)

A Figura 3 mostra um exemplo de uma matriz de contexto de palavra entre substantivos e seus verbos vizinhos. As linhas representam as palavras de interesse (vetores de palavras) e as colunas representam o contexto (vetores de contexto). Os valores nas células são as frequências com que as palavras ocorreram juntas. A partir desta representação, partindo das contagens nas linhas, já podemos observar que bicicleta e carro compartilham contextos semelhantes, da mesma forma que cão e leão (Agt-Rickauer, 2019:50).

Qual é a melhor abordagem?

Em 2013, três pesquisadores (Chiticariu et al, 2013), cujas atuações localizam-se na fronteira entre academia e indústria, realizaram uma pesquisa em que examinam os artigos dos principais periódicos científicos da área de linguística computacional dos dez anos anteriores e confrontam com dados obtidos a partir de

[14] Firth, 1957 apud Agt-Rickauer, 2019.

uma extensa análise de produtos disponíveis no mercado para fins de IE. O que eles descobriram foi que a grande maioria dos sistemas desenvolvidos comercialmente eram baseados em regras, incluindo aí sistemas de grandes companhias como IBM, SAP e Microsoft. Apenas 1/3 dos 54 produtos examinados dependia inteiramente de ML.

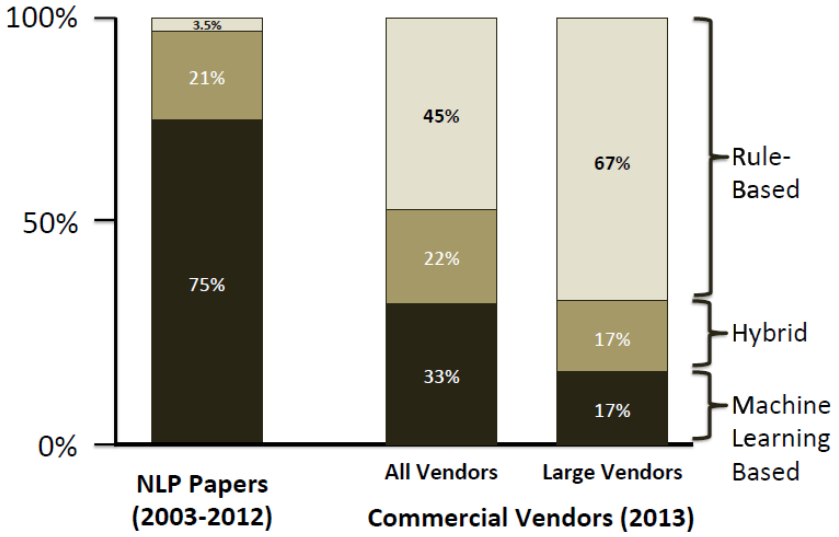

Figura 4 - Comparação entre as abordagens adotadas em sistemas de IE (tarefa NER) academia x indústria (extraído de Chiticariu, 2013)

Para os autores, essa desconexão surge da diferença em como as duas comunidades medem os custos e benefícios da extração de informação. Enquanto a academia avalia o desempenho dos sistemas em termos de precisão e abrangência – principalmente em contextos de competições –, a indústria tende a valorizar mais a capacidade de incorporação de conhecimento de domínio e ataque aos problemas de negócios específicos, ou seja, soluções que sejam acessíveis e factíveis para aplicação, depuração, manutenção e adaptação face a requisitos de mudanças. Um exemplo seria o reconhecimento de nomes completos de pessoas sem as formas de tratamento comuns (Sr., Sra. etc.): em um sistema baseado em regras uma expressão regular simples seria o suficiente, mas no aprendizado de máquina um retreino completo se faz necessário.

Os autores apontam que dentre os argumentos que pesquisadores em PLN costumam usar para depreciar a abordagem dos padrões é que escrever regras é uma tarefa tediosa e demorada, e não se vê aí um verdadeiro "problema de pesquisa" (Chiticariu et al, 2013:830). No entanto, não se pode ignorar (como a literatura em

PLN parece fazer, segundo os autores) que as próprias tarefas de ML são complexas e também consomem muito tempo, sendo necessário: i) definir o problema alvo em termos matemáticos, ii) medir as escolhas dentre os modelos de aprendizagem, iii) levar a cabo a engenharia do processo com base em uma sólida compreensão do modelo escolhido, e iv) reunir grandes quantidades de dados etiquetados, muitas vezes usando outro processo de automação inteligente. Além de demandar um alto grau de qualificação, a opacidade inerente aos sistemas de ML torna a adoção da solução algo difícil, arriscado, e pouco prático (Chiticariu et al, 2013).

Para os autores, uma agenda de pesquisa voltada para sistemas baseados em regras pode abrigar inúmeros tópicos. Um deles poderia ser a busca por uma linguagem padrão para expressar as regras, que com sua expressividade permita ao especialista se preocupar com *o que* o programa deve obter e não com o *como*. Em paralelo, investigações sobre o modelo de dados ideal para capturar texto, anotações sobre texto e suas propriedades, focando em questões sistêmicas, tais como representação de dados e avaliação de regra via otimização automática de desempenho. Por fim, com esse conjunto de construtos e uma linguagem padrão de regras, motivar o desenvolvimento de novas soluções híbridas de IE junto com a tecnologia de ML (Chiticariu et al, 2013).

3
Trabalhos relacionados

Neste capítulo são visitadas iniciativas realizadas no campo da extração de informações a partir de abordagens baseadas em padrões textuais, puras ou combinadas com aprendizado de máquina.

Em Hearst (1992) a técnica é investigada a partir de relações de hiponímia – que são as relações do tipo *is-a* –, conforme o exemplo abaixo:

$$NP_0 \text{ such as } \{NP_1, NP_2 ..., (and \mid or)\} NP_n$$

Na expressão, NP é qualquer sintagma nominal que a frase contém. Aplicando este padrão em um corpus a ideia é obter exemplos como “countries such as Brazil, Italy and Norway”, que por sua vez inferirá três instâncias de países: is-a(Brazil, country), is-a(Italy, country), is-a(Norway, country) .

O processo percorrido por Hearst para a descoberta de novos padrões é descrito da seguinte forma:

1) Decidir a relação lexical de interesse (por ex. hiponímia);
2) Reunir uma lista de termos para os quais essa relação é válida (por exemplo, "Brazil-country", “car-vehicle”);
3) Procurar no corpus por sentenças onde esses termos ocorrem sintaticamente próximos e guardar estes trechos;
4) Encontrar as semelhanças existentes nestas sentenças e hipotetizar quais contextos comuns (padrões) indicariam a relação de interesse;
5) Quando o novo padrão tiver sido positivamente identificado, usá-lo para encontrar mais instâncias da relação de interesse, e então voltar para o passo 2.

Para testar sua proposta, Hearst aplicou os seus padrões textuais sobre um corpus jornalístico contendo seis meses do New York Times e avaliou os resultados

comparando as relações extraídas com os dados da Wordnet[15]. Seus dois artigos (1992, 1998) tornaram-se seminais e inspiraram muitos trabalhos na área desde então.

Na sétima edição do MUC, o trabalho de Mikheev e colegas (1998), obteve bons resultados em um sistema de NER utilizando regras de contexto como as apresentadas abaixo:

Regra de contexto	Atribuição	Exemplo
Xxxx+ is a? JJ* PROF	PERS	Yuri Gromov is a former diretor
PERSON-NAME is a? JJ* REL	PERS	John White is beloved brother
Xxxx+, a JJ* PROF,	PERS	White, a retired 64alidos,
Xxxx+, whose REL	PERS	Nunberg, whose stepfather
Xxxx+, DD+,	PERS	White, 33,
PROF of/at/with Xxxx+	ORG	Director of Trinity Motors
In/at Xxxx+	LOC	In Washington
Xxxx+ área	LOC	Beribidjan área

Tabela 3 - Regras de contexto para tarefa de NER (Mikheev et al, 1998).

As etiquetas podem ser lidas da seguinte forma: “Xxxx+ é uma sequência de palavra iniciada em maiúscula; DD é um dígito; PROF é uma profissão; REL é um parentesco; JJ* é uma sequência de zero ou mais adjetivos; LOC é um local; PERSON-NAME é um nome próprio de pessoa”. A partir dessas regras *slots* são preenchidos para identificar entidades do tipo pessoas, organizações e locais. A medida-F alcançada pelo sistema foi de 93,39%, a mais alta dentre os competidores do MUC-7.

No ConLL de 2003, Florian e colegas (2003) apresentam um modelo que combina abordagem híbrida para a identificação e classificação de entidades nomeadas. Um conjunto de *features* é acessado ao se examinar uma palavra em contexto, como por exemplo “os lemas das cinco palavras à sua esquerda e à direita”, “as marcações gramaticais (POS tag) e das palavras próximas”, “prefixos e sufixos”, “morfologia (primeira letra em maiúscula, número de dígitos, todas as letras em maiúsculas, etc)”. Além disso, os autores utilizam listas lexicais contendo cerca de 50 mil cidades, 80 mil nomes próprios e 3,5 mil organizações. Um dos testes era para avaliar a eficácia do emprego destes pequenos dicionários, e a conclusão que chegaram ao final foi que eles garantiram uma redução de erro (*F*-

[15] https://wordnet.princeton.edu/

measure error) de cerca de 15 a 21% na tarefa de reconhecimento de entidades com uso destas listas.

Girju e colegas (2003) buscaram extrair relações de meronímia, ou seja, relações do tipo *part-of* a partir de um corpus constituído pelo jornal Los Angeles e uma parte do English Brown Corpus. Esta relação, segundo os autores, tem variadas formas de expressão, e por consequência, um número grande de estruturas léxico-sintáticas possíveis. Algumas são inequívocas, como por exemplo, "The substance consists of two ingredientes" e "Iceland is a member of NATO", porém outras podem ser ambíguas: "He is part of the game".

Foram encontrados os padrões:

NP of PP
NP's PP
NP verb NP

PP é qualquer sintagma preposicional que a frase contém. A estratégia adotada para identificar as formas lexicais que expressam essa relação parte-todo, foi primeiramente buscar na Wordnet pares-semente onde essa relação era estabelecida, e em seguida, localizar nos corpora as sentenças em que esses pares compareciam. Cerca de 10,000 sentenças foram selecionadas e manualmente inspecionadas, permanecendo apenas aquelas onde os pares conectavam-se pela relação de meronímia.

Para tratar casos de ambiguidade em que o padrão recaía em outro tipo de relação, os autores propuseram definir restrições semânticas sobre os argumentos, da seguinte forma: primeiro, um conjunto de treinamento é criado manualmente contendo tuplas representando uma possível relação de meronímia, classificadas como corretas ou incorretas. Por exemplo, <aria; opera; yes> indicaria que "aria" é parte de uma "opera". Em seguida, esses exemplos são generalizados substituindo-se os termos pelos seus sentidos da Wordnet. Assim, <aria, opera, yes> torna-se <aria#1, entity#1; opera#1, abstraction#6; yes>, onde o conceito "aria" pertence ao conceito "entity" e o conceito "opera" é parte do conceito "abstraction", na hierarquia definida pela Wordnet. A partir desse conjunto de restrições os algoritmos são então treinados para resolverem casos ambíguos da relação.

O estudo de Freitas (2007) é fortemente influenciado por Hearst e outros trabalhos que desenvolveram algumas técnicas de regras de associação (Morin & Jacquemin, 2004; Cederberg & Widdows, 2003). O objetivo foi investigar a possibilidade de elaboração automática de uma ontologia sem a determinação prévia das categorias que a compõem. A autora adapta os padrões de Hearst para a língua portuguesa e propõe alguns novos a partir da observação do corpus, produzindo expressões regulares para extração de relações de hiperonímia e correferência. Por exemplo, a sentença "e nele existe uma substância chamada benzopireno", o seguinte padrão foi descoberto:

SN HHiper chamado/s/a/as (de) SN Hipo

Devolvendo, para o caso acima, algo como é-um(benzopireno, substância). No total seis padrões foram aplicados em um corpus no domínio da saúde e as relações obtidas avaliadas por humanos de acordo com uma pontuação pré-estabelecida de erros e acertos. A pesquisa mostrou que a metodologia pode ser de grande valia para investigações lexicográficas e linguísticas, além de guardar grande potencial para a descoberta do maior número possível de relações entre entidades, através do levantamento dos verbos que aparecem no corpus (Freitas, 2007).

Giovannetti e colegas (2008) descrevem uma metodologia híbrida para extração de relações semânticas. Por um lado, padrões léxico sintáticos genéricos são aplicados a um corpus anotado para detectar um primeiro conjunto de pares de palavras que co-ocorrem, possivelmente envolvidos em relações sintagmáticas (por exemplo, *steer* e *car* na sentença "*steer is part of the car*"). Por outro lado, um sistema estatístico não supervisionado de associação é usado para obter um segundo conjunto de pares de termos com similaridade distribucional, que parecem ocorrer em contextos similares, sendo possivelmente envolvidos em relações paradigmáticas (por exemplo, os termos *car* e *motorcycle* nas sentenças "*I drive my car*" e "*Bob drives his motorcycle*"). A abordagem visa o aprendizado de informação ontológica a partir do filtro de relações candidatas obtidas através dos padrões genéricos e pelas etiquetas das relações anônimas atribuídas pelo sistema estatístico.

Daniel Santos e colegas (2010), apresentam um sistema de extração de relações familiares para a língua portuguesa, criando um conjunto de *features* específico para este tipo de relação. A técnica adotada é apoiada em regras aplicadas em um corpus anotado morfossintaticamente. Taba (2013) utiliza abordagem híbrida, além de uma base de conhecimento aberta chamada "Senso Comum", contendo fatos e crenças compartilhados por uma comunidade. Estes fatos, traduzidos em triplas semânticas, atuam como sementes para extrair mais instâncias de relações. Makarov (2018) utiliza padrões para derivar construções sintáticas sobre conflitos e protestos políticos a partir de textos de notícias. Uma supervisão fraca é usada para identificar possíveis candidatos e aprender sobre as distribuições semânticas.

4
Dicionário Histórico-Biográfico Brasileiro

Nas seções seguintes apresentamos as características do DHBB, sua importância como referência para estudos sobre a história política contemporânea do Brasil, e as possibilidades de pesquisa como corpus anotado.

4.1
Sobre a obra

Em 1984, quando o Dicionário Histórico-Biográfico Brasileiro foi lançado, o acesso à Internet não era algo generalizado e fazer pesquisa sobre temas da história política contemporânea do Brasil significava – para além das leituras obrigatórias – sair em campo e consultar arquivos privados e públicos, jornais e periódicos, fontes oficiais e outros registros orais e escritos, nem sempre de fácil acesso. Ao ter as informações organizadas e sistematizadas, o DHBB tornou-se em pouco tempo referência para historiadores, cientistas políticos, jornalistas e demais interessados na política brasileira do período de 1930 até os dias atuais.

Em sua primeira edição, a obra continha quatro grandes volumes impressos com cerca de 4.500 verbetes biográficos e temáticos. Dezessete anos depois, em 2001, esse conteúdo foi atualizado e acrescido de mais um volume, somando 6.620 entradas. Foi em 2010 que ela ganhou uma versão digital inteiramente disponível online, e desde então tem sido regularmente revisada e atualizada, contabilizando neste ano de 2019 mais de 7.500 entradas, 90% sendo de caráter biográfico. Para uma ideia do que encontramos no DHBB, a tabela abaixo traz alguns dos principais cargos ou papéis associados aos seus personagens[16] (e, claro, um personagem pode exercer mais de um papel ao longo de sua vida).

Cargo	Total de ocorrências
Presidentes da República	26
Ministros	894
Senadores	742
Deputados Federais	4.314

[16] A tabela completa com todos os cargos encontra-se no Anexo 4.

Militares	800
Jornalistas	219
Religiosos	73

Tabela 4 - Amostra do que é possível encontrar no DHBB

O DHBB possui critérios específicos para decidir se uma personalidade, evento, conceito ou instituição deve se tornar um verbete. Esses critérios, embora não rígidos em determinados aspectos, se destinam a preservar os objetivos de dar a maior abrangência e relevância possível na cobertura de um universo tão amplo e multifacetado como é a história política contemporânea do Brasil. Para biografias, todo aquele que tenha ocupado posição relevante ao nível federal da administração pública do país foi incluída, ficando de fora a maior porção do mundo político regional e municipal. Uma importante amostra da sociedade civil brasileira é representada com a presença de certos presidentes de organizações, entidades e empresas privadas. Os principais líderes de rebeliões e protagonistas que detinham posições informais de poder também foram incluídos.

Uma característica importante do DHBB e que pode se apresentar como providencial para fins de extração, é a uniformidade da escrita dos verbetes. Desde o uso do tempo verbal, passando pela normalização dos nomes até a ordenação do conteúdo, tal padronização pode vir a se traduzir em mais abrangência no que diz respeito às regras a serem utilizadas para a extração, em especial na identificação das relações semânticas entre as entidades: “diferentes relações podem ser expressas utilizando um pequeno número de padrões léxico-sintáticos” (Hearst, 1992).

4.2
Fonte para estudos prosopográficos

No final dos anos 1980, um estudo de análise estatística orientado pelo brasilianista Michael Conniff (2003, 2006) com uma amostra de 7% dos verbetes do DHBB (cerca de 250 biografias à época), permitiu a localização de mudanças importantes relativas à origem, idade, formação e classe social da elite política brasileira. Com a ajuda de assistentes humanos, foi realizada uma seleção dos indivíduos que ocuparam cargos importantes no poder Executivo em um determinado período, e a partir de uma leitura individual e atenta dos textos,

informações de interesse foram extraídas manualmente e sistematizadas em uma base SPSS[17].

Com isso, a equipe foi capaz de mapear vários características e mudanças sofridas por esta elite. Por exemplo, no início do século XX a maioria era formada por homens de meia-idade que tipicamente entravam na política como uma segunda ou terceira carreira. Com o tempo esse perfil foi se tornando cada vez mais jovem; na média aqueles que nasceram antes de 1900 começavam na carreira política aos 55 anos, os que nasceram entre 1901 e 1920 começavam aos 37, e os que nasceram após 1921 começavam aos 32 anos. Na educação formal, o mais comum era ter diploma em Direito (44%) seguido de uma carreira militar. Engenheiros e médicos viriam em seguida com 12% e 5% respectivamente. A mudança mais marcante identificada por Conniff nesta área é o declínio da carreira militar ao longo dos anos: enquanto para aqueles nascidos antes de 1920, 37% tinha essa formação, para os nascidos após 1920, isso representava apenas 10%.

Hoje, se um pesquisador a consultar o DHBB está interessado em informações desse tipo, ele precisará ler todos os verbetes que tem a ver com o assunto. Qual é o caminho mais frequentemente trilhado para se chegar à presidência? Qual o perfil acadêmico dos membros do congresso nacional das últimas legislaturas? Qual a idade média de ingresso de um juiz na Suprema Corte Federal? O que dizer sobre os vínculos familiares entre políticos?

Quando historiadores e cientistas de dados se reúnem em cooperação, são os primeiros que, além de fornecer o corpus e os dados, trazem as perguntas e problemas que gostariam de ver resolvidos. Em conjunto devem descobrir o potencial existente nesses dados e pensar nas possibilidades de análise, seja utilizando técnicas de IE, mineração de texto, modelagem de tópicos ou outras. Em termos de humanidades digitais, podemos dizer que se trata de querer fazer uma leitura distante para a História.

[17] Statistical Package for Social Sciences

4.3
O corpus DHBB

Enciclopédias ocupam um papel importante na história cultural, visto que têm a intenção de encapsular de forma sistemática a totalidade do conhecimento humano, ou de um campo específico, em seus textos. No entanto, o problema com enciclopédias para o acadêmico é a vasta quantidade de informações e o número proporcionalmente grande de tópicos que elas contêm, sugerindo que métodos das humanidades digitais podem ser úteis para auxiliar historiadores, linguistas e outros estudiosos culturais a lidar com esse grande volume de textos.

Algumas iniciativas de pesquisa em PLN tendo o DHBB como objeto empírico foram conduzidas nos últimos anos, com objetivos diversos, tais como, desenvolvimento de recursos linguísticos, treinamento de analisadores ou apenas estudos de fenômenos da língua. Em Higuchi e colegas (2018), os autores compararam e avaliaram as anotações geradas por dois diferentes parsers (o PALAVRAS, baseado em regras e gramática restritiva, e o UD-Pipe, que utiliza aprendizado de máquina e foi treinado com um corpus golden em português), observando especialmente para as relações sintáticas apositivas, fenômeno linguístico que induz diferentes relações semânticas entre entidades (Higuchi et al, 2018). Já no artigo Distant reading Brazilian politics, Higuchi e colegas (2019), relatam o trabalho em andamento de enriquecimento semântico do corpus visando a extração de informações, focando em particular para as tarefas de desambiguação de nomes próprios e incorporação de anotação semântica do domínio das relações familiares. Em Freitas e Souza (2021), os autores apresentam estudos descritivos e computacionais relacionados ao fenômeno de sujeito oculto nas sentenças e utilizam o DHBB como um dos três corpora para análise, levantamento dos desafios e possíveis caminhos para endereçar a questão, que tanto impacta nas tarefas de PLN.

Em 2018, o DHBB foi integrado ao acervo do AC/DC[18] e atualmente, na versão 7.3, é constituído de 457.101 frases, quase 16 milhões de tokens e cerca de 14 milhões de palavras.

[18] https://www.linguateca.pt/acesso/corpus.php?corpus=DHBB

4.3.1
Sobre as entidades e relações

Com o cuidado de não perder de vista o domínio sobre o qual estamos tratando – a história política contemporânea do Brasil –, definimos um primeiro conjunto de categorias. Elas foram se delineando a partir da leitura dos verbetes e de acordo com as questões de competência da pesquisa. Por exemplo: *quem* é a pessoa P?, *onde* e *quando* a pessoa P nasceu?, em que *instituições* a pessoa P estudou?, com quais *organizações* a pessoa P se vinculou? em qual(is) *evento*(s) a pessoa P se envolveu?

São estas as categorias resultantes dessa análise preliminar: pessoas, organizações, formulações políticas, eventos/processos, tempo, espaço e documentos.

Na versão atual do DHBB foram identificadas cerca de 1,6 milhão de nomes próprios, sendo mais de 120 mil distintos:

Verbetes	Biográficos	6.717
	Temáticos	968
Sentenças		457 mil
palavras (formas)		14 milhões
palavras distintas		120 mil
nomes próprios		1,6 milhão
nomes próprios distintos		121 mil
nomes de pessoas		49,7 mil
nomes de organizações		41 mil
nomes de lugares		5,4 mil
eventos		3,9 mil
formulações políticas		560

Tabela 5 - Visão geral do DHBB

A maioria, quase 50 mil, se referem a pessoas. Podemos notar que a classe criada especialmente para o DHBB, "formulações políticas" foi a que teve menos ocorrências, e isso acontece porque estas entidades vieram de listas compiladas manualmente (ver seção 5.3.3). Ainda não foi feita uma revisão para verificar o

índice de acerto do analisador na atribuição dos outros nomes próprios, o que deverá acontecer no futuro.

Apresentamos a seguir uma tabela de equivalência das classes que definimos como relevantes para o DHBB com as classes do PALAVRAS, analisador morfossintático com o qual estamos trabalhando:

Entidades do DHBB	CLASSES EQUIVALENTES NO PALAVRAS[19]
PESSOA Nesta classe incluímos todas as entidades que possam recuperar instâncias de indivíduos (*Getúlio Vargas, Lula, presidente...*)	**<hum> -** Nomes de pessoa **<hfam>** Familia **<official> , <member>** **<Hprof> -** Profissional humano **<Htit>** - nome de título *(rei, rainha, duque, excelência…)* **<HHsoc>** - Grupo socialmente definido (*nobreza…*)
ORGANIZACAO Esta classe diz respeito à entidades que podem ser do tipo empresa privada ou governamental (*Companhia de Cimento Vale do Paraíba, Petrobras)*; instituição de ensino (*Ginásio Diocesano de São Paulo, Faculdade de Ciências Jurídicas e Sociais)*; órgãos de imprensa (*Correio da Bahia, TV Aratu, Tribuna da Bahia, Rede Manchete)*; organização político-administrativa (*Ministério das Minas e Energia, Câmara, Assembléia Nacional Constituinte, Comissão Permanente de Obras Públicas, Partido Democrático Social)*; instância jurídica (*Supremo Tribunal Federal, Superior Tribunal de Justiça);* além de todo grupo de indivíduos, formalizado ou não, que tenha um nome cristalizado ou convencional (*comissão parlamentar de inquérito, analistas políticos, revolucionários constitucionalistas*).	**<org>** Nomes de organização **<inst>** Instituições em geral **<media>** - Empresa de mídia, imprensa **<party>** - Partido político **<HH> -** Grupo de humanos (*grupo* guerrilheiro*, comissão parlamentar…*) **<HHorg>** - Grupo organizado
FORMULACAO_POLITICA Nesta classe entram tanto planos, programas, projetos, reformas e alianças (como *Projeto de Irrigação Formoso, Aliança para o Progresso, Plano Collor, Programa de Estabilização Monetária, Reformas de Base*] quanto acordos, tratados e cooperações (como *Acordo Nuclear Brasil-Alemanha, Acordo Comercial Brasil-Estados Unidos, Acordo MEC-USAID, Tratado de Cooperação Amazônica*), além de leis, decretos, códigos e constituições.	**<tit>** - títulos de trabalhos em geral, livros e periódicos **<conv> Collective prototypes** - Para leis **<prop> -** acordo, lei... **<sem-c> Semantic product prototypes -** Produtos de cognição (*esquema, plano, acordo…*)
EVENTO Evento denota uma ocorrência pública marcante sob algum aspecto (como uma reunião, manifestação, desfile, deposição, revolta, golpe, passeata, atentado).	**<occ> -** Para nomes de ocasiões, eventos complexos (jantar, desfile, *Revolução Russa, I Guerra Mundial…*) **<sit> State-of-affairs prototypes -** Situação psicológica ou estado físico (*crise, ilegalidade…*)

[19] http://beta.visl.sdu.dk/visl2/semantic_prototypes_overview.pdf

	<event> Time and event prototypes - Evento sem controle (*milagre, explosão, acidente…*) **<act> Action prototyes** - Ações do tipo *abdicação, unificação, mobilização...* **<process> Perception prototypes** - Processos do tipo *estabilização, balcanização…*
LOCAL A noção de local aqui é abordada a partir de dois tipos específicos: [GEOP] **GEOPOLÍTICO**, ou seja, espaços delimitados política e/ou administrativamente (*Distrito Federal, Rio de Janeiro, Nordeste*); e [CONST] **CONSTRUÍDO**, como instalações, monumentos, construções **(***palácio de Ondina, palácio Guanabara, porto de Alcântara, embaixadas)*.	**<civ>** - Para cidade, país, município (com administração) **<top>** - Para espaço geográfico **<build>** - Para prédios, construções... **<Lciv> -** Civitas, cidade, país, município **<Lh> -** Espaço humano funcional (anfiteatro, estádio, elevador…) **<Ltop>** - Espaço geográfico, natural
DOCUMENTO Distinguimos os seguintes tipos: [PRIV] **PRODUCAO_PRIVADA**, demandado principalmente para as obras produzidas pelos biografados, e também cartas, anotações e diários; e [PRESS] **IMPRENSA** para jornais e revistas.	**<sem-r> Semantic Product prototype -** Coisa de leitura (*jornal, revista..*)
TEMPO Esta categoria abrange os seguintes tipos: [DAT] **DATA**, para abranger dia, mês e ano; [PER] **PERÍODO**, para qualquer menção como (*1927 a 1929, 1927-1929);* [DUR] **DURACAO**, para expressões como (*seis meses, dois anos, 5 horas).*	**<dur>** **<temp> -** Conceitos temporais que apontam no tempo **<per>** - Período do tempo

Tabela 6 - Classes de entidades relevantes para o DHBB e mapeamento com as classes existentes no PALAVRAS

Em Santos e colegas (2020), encontramos uma descrição detalhada da lógica e do funcionamento do sistema PALAVRAS-NER e a sequência de comandos do analisador utilizada para identificar e classificar as entidades mencionadas em um corpus.

Sobre relações semânticas úteis para o DHBB

Que relações semânticas consideramos úteis para o DHBB? Por ser um material extremamente rico contendo diversas informações sobre personagens e cenários políticos do Brasil, identificamos abaixo um conjunto de relações fundamentais para futuros trabalhos de extração de informação automática. Obviamente, é uma lista preliminar que estenderá conforme as necessidades de

informação sejam endereçadas ao corpus. Ela é inspirada nos esquemas definidos em alguns fóruns de competição de IE, como os do Segundo HAREM, na avaliação do ReRelEM[20]

> Ident – para correferência e entre itens lexicais que querem dizer a mesma coisa ou relacionam-se de alguma forma
>
> papel – liga uma pessoa (que denota um papel) a outra pessoa (indivíduo ou grupo)
>
> participante – liga uma pessoa ou organização a um evento
>
> local_nascimento – vincula uma pessoa ao local de seu nascimento
>
> local_falecimento – vincula uma pessoa ao local de sua morte
>
> data_nascimento – vincula uma pessoa à data de seu nascimento
>
> data_falecimento – vincula uma pessoa à data de seu falecimento
>
> data_de – liga uma data a um acontecimento
>
> ocorre_em – liga alguma coisa (evento, organização) a um lugar
>
> vinculo_inst – liga uma organização a uma pessoa
>
> vinculo_familiar – liga uma pessoa à outra, não importa o grau de parentesco.

A relação vínculo-familiar, como veremos no próximo capítulo, será usada para extrairmos informações do DHBB, como parte do nosso trabalho empírico.

4.3.2 Alguns desafios de PLN

O desafio de transformar o DHBB em um corpus anotado, a partir do qual podemos extrair informações automaticamente, não é pequeno. Além das limitações impostas pelas ferramentas, existem as dificuldades que a própria linguagem natural e seus falantes se encarregam de trazer ao processo, e que requer mais do que apenas compreender as sentenças. A seguir descrevemos alguns desses desafios, sem, no entanto, termos a pretensão de tentar resolvê-los nesta pesquisa.

Construção sintática

[20] https://www.linguateca.pt/harem/avaliacao/?tipo=rerelem

A linguagem natural permite que nos expressemos de forma diversa para descrever uma mesma situação. Por exemplo:

> "O governador Leonel Brizola é cunhado do presidente João Goulart"
>
> "João Goulart, presidente, é cunhado de Leonel Brizola, governador"
>
> "O cunhado do presidente João Goulart, o governador Leonel Brizola..."
>
> "João Goulart, presidente, e Leonel Brizola, governador, são cunhados..."

As sentenças acima, criadas para ilustração, estão todas sintaticamente corretas e mostram a relação existente entre João Goulart e Leonel Brizola, com seus respectivos papéis. Com a nossa capacidade cognitiva conseguimos extrair de cada uma delas as mesmas informações, independente da ordem dos seus constituintes e da construção sintática utilizada. Mas transformar todo este entendimento em estruturas formais legíveis por máquina não é um processo simples.

Ademais, a riqueza de construções possíveis para expressar determinado conhecimento impacta diretamente na definição do número de padrões que serão necessários para extrair informação do corpus quando utilizamos a abordagem por regras.

Multiword expression (MWE)

Expressões multivocabulares ou multipalavras representam um grande desafio para sistemas de PLN. Villavicencio e colegas (2010) as definem como combinações de palavras que apresentam idiossincrasias lexicais, sintáticas, semânticas ou estatísticas (Villavicencio et al, 2010:18). Fazem parte destas combinações as expressões idiomáticas ("perder a linha", "fazer vista grossa"), os verbos de suporte ("dar uma palestra"), os compostos nominais ("deputado federal") e os nomes próprios ("Rio de Janeiro", "João Goulart").

MWEs estão muito presentes na língua. Elas podem variar de 30% a 45% do inglês falado (Biber et al, 1999 apud Villavicencio et al, 2010) e correspondem a mais de 40% das entradas na Wordnet (versão 1.7, segundo Sag et al, 2002). Na língua portuguesa e no domínio do DHBB, em particular, não é diferente.

Encontramos muitos exemplos, como "Dicionário Histórico-Biográfico Brasileiro", "Comissão Parlamentar de Inquérito" e "Golpe de 1964", e todas elas somente terão utilidade no processamento se forem capturadas como itens lexicais não circunscritos à morfologia, ou seja, se forem capturadas em blocos únicos de significação.

A eficácia no tratamento das MWEs vai depender das técnicas utilizadas pelos analisadores. Na literatura encontramos discussões sobre algumas estratégias adotadas para lidar com o desafio, que envolvem estatísticas de associação entre palavras, alinhamento sentencial com corpus paralelo (Villavicencio et al, 2010), regras combinatórias restritas, gramáticas formais (Sag et al, 2002), além de proposta de pipeline combinando várias técnicas para a aquisição de MWEs (Ramisch, 2012).

Tratar adequadamente estas expressões também se aproxima da questão das entidades encaixadas, conforme vimos na seção anterior. Em muitos casos será preciso consertar segmentações erradas ou juntar palavras compostas através de regras manuais, conforme veremos na seção 5.3.4.

Ambiguidade e vagueza

No DHBB encontramos muitos casos de pessoas sendo mencionadas de várias formas, pelos seus nomes abreviados, por apelidos e até mesmo com erros de grafia. Há casos em que homônimos só são resolvidos quando associados aos seus nomes completos de família. Esse trabalho de identificação é conhecido em inglês como *entity grounding*[21] ou *entity disambiguation*[22], onde criamos correspondências entre as formas, atribuindo a elas uma identificação única. Sem isso, corremos o risco de perder informação. Na seção 5.3.4, apresentamos a estratégia adotada para minimizar este problema no corpus.

Casos de metonímia, onde um nome originalmente usado para denotar certa entidade é usado como substituto para outras entidades, são recorrentes no processo de reconhecimento das entidades mencionadas. Como Markert e Nissim (2002, apud Santos, 2007) apontaram, padrões metonímicos podem ocorrer associados a

[21] https://en.wikipedia.org/wiki/Symbol_grounding_problem
[22] https://en.wikipedia.org/wiki/Entity_linking

todo tipo de entidade, como por exemplo, *place-for-event*, *place-for-people*, *place-for-product*, etc. Assim, Palácio do Planalto pode ser utilizado para denotar construção ou governo; Folha de S. Paulo, para empresa ou publicação; Brasil, para denotar lugar, instituição ou povo; e assim por diante. A vagueza ocorre quando o uso de uma palavra gera casos duvidosos de aplicação a certos seres ou situações, ou seja, uma circunstância na qual mais de uma interpretação poderá ser atribuída a um mesmo item. Identificar a relação de sentido correta nestes casos é uma habilidade cognitiva difícil de reproduzir automaticamente. Haverá diferentes casos em que um utilizador estará interessado em diferentes vertentes de um mesmo conceito (Santos, 2007:49).

Sujeito oculto

Sujeito oculto (ou sujeito elíptico) é aquele que não está explícito na oração, mas pode ser determinado pela flexão número-pessoa do verbo, ou por sua presença em alguma oração antecedente. Por exemplo, no trecho retirado do verbete de Fernando Collor de Melo, "Em 1975, *casou-se* com Celi Elizabeth Júlia (Lilibeth) Monteiro de Carvalho...", o sujeito que se casou com Lilibeth de Carvalho é o próprio Fernando Collor, que foi mencionado em outro parágrafo muito anterior a esse trecho.

No DHBB, o que ocorre bastante é que o nome do biografado é mencionado apenas no início do primeiro parágrafo do seu verbete e daí em diante o sujeito fica implícito nos verbos flexionados. Na extração automática de informações, as máquinas têm ainda muitas limitações para tratar esses fenômenos, e esse desafio impacta diretamente na qualidade das extrações, que depende de triplas <entidade, relação, entidade> para derivar conhecimento útil.

Não é objetivo da pesquisa tentar resolver este problema no momento, apenas apresentar algumas considerações envolvendo a questão e mensurar o impacto deste fenômeno da tarefa de IE. Para isso recorremos aos trabalhos de Martins e Freitas (2019) e Freitas e Souza (2019).

No primeiro trabalho Martins e Freitas utilizaram uma amostra de seis verbetes biográficos para investigar se os verbos sem sujeito presentes nos textos dizem respeito ao próprio biografado ou a uma outra pessoa, ou ainda se se trata de sujeito indeterminado. Como metodologia, as autoras processaram o corpus com o UDPipe (Straka e Strakova, 2016) para identificar os verbos, considerando todas as formas conjugadas, os gerúndios, os infinitivos e os particípios nesta análise. Foram encontrados 2.920 verbos, sendo que destes, 666 são verbos sem sujeito. Checando manualmente estas ocorrências, puderam concluir que a grande maioria, 95%, referem-se mesmo ao verbetado:

> Durante o governo de Itamar Franco, (*ÁLVARODIAS*) **foi sondado** para (*ÁLVARODIAS*) **ocupar** diversos postos, dentre os quais o de ministro da Indústria e Comércio, não (*ÁLVARODIAS*) **chegando**, no entanto, a (*ÁLVARODIAS*) **assumir** nenhum cargo público, (*ÁLVARODIAS*) **preferindo** (*ÁLVARODIAS*) **dedica**r-se à organização do novo partido. (Verbete ÁLVARODIAS)

Apenas 2% dizem respeito a uma terceira pessoa mencionada anteriormente na sentença:

> Antigo aliado de Fernando Henrique Cardoso, **Antônio Carlos Magalhães** havia-se rebelado contra o apoio do governo à candidatura de seu adversário Jader Barbalho (PMDB-PA) na eleição para a presidência do Senado em fevereiro de 2001. (*OUTRO*) **Acusado** de quebra de decoro parlamentar pela violação do painel, (*OUTRO*) **renunciaria** ainda em maio ao mandato de senador. (Verbete CIRO GOMES)

E 4% são casos de sujeito indeterminado:

> Naquele mesmo mês, (*JAIRBOLSONARO*) referindo-se à denúncia do ex-padre José Antônio Monteiro de que o diretor-geral da Polícia Federal João Batista Campelo o havia torturado, o que resultou no afastamento deste último, (*JAIRBOLSONARO*) afirmou ser isso "que dá (*INDET*) **torturar** e não (*INDET*) **matar**". (Verbete JAIR BOLSONARO)

Já no segundo trabalho, Freitas e Souza (2019) investigam a presença do sujeito oculto em três corpora dos gêneros jornalístico, literário e enciclopédico,

este último representado pelo DHBB. Segundo os autores, o corpus DHBB é o que apresenta a maior quantidade de omissões de sujeito, presumivelmente pela natureza de seu conteúdo: “verbetes biográficos ou temáticos, nos quais o tema/ foco da frase dificilmente se altera, e, por isso, a omissão é o recurso estilístico utilizado para deixar o texto não repetitivo” (Freitas e Souza, 2019:18).

A omissão do sujeito ficou em 46% do total de orações existentes no DHBB contra 41% do corpus literário e 24% do corpus jornalístico. Para testar o quanto essa omissão prejudica as análises do processamento automático, os autores conduziram um experimento de PLN no corpus jornalístico (o único revisto) primeiro reintroduzindo os sujeitos omitidos e em seguida treinando um modelo de aprendizagem de máquina para avaliar os resultados.

O desenvolvimento de estratégias para a reconstituição correta do sujeito é apresentado pelos autores a partir de algumas heurísticas e regras aplicadas, um exercício desafiador, mas que teve um índice de acerto acima de 80% para cada estratégia, o que é bastante promissor.

5
Consolidação da metodologia

Neste capítulo apresentamos as etapas percorridas. Algumas questões podem vir a ser introduzidas, pois decorrem dos resultados e da manipulação dos dados tratados. A figura representa simplificadamente o *workflow* do processo.

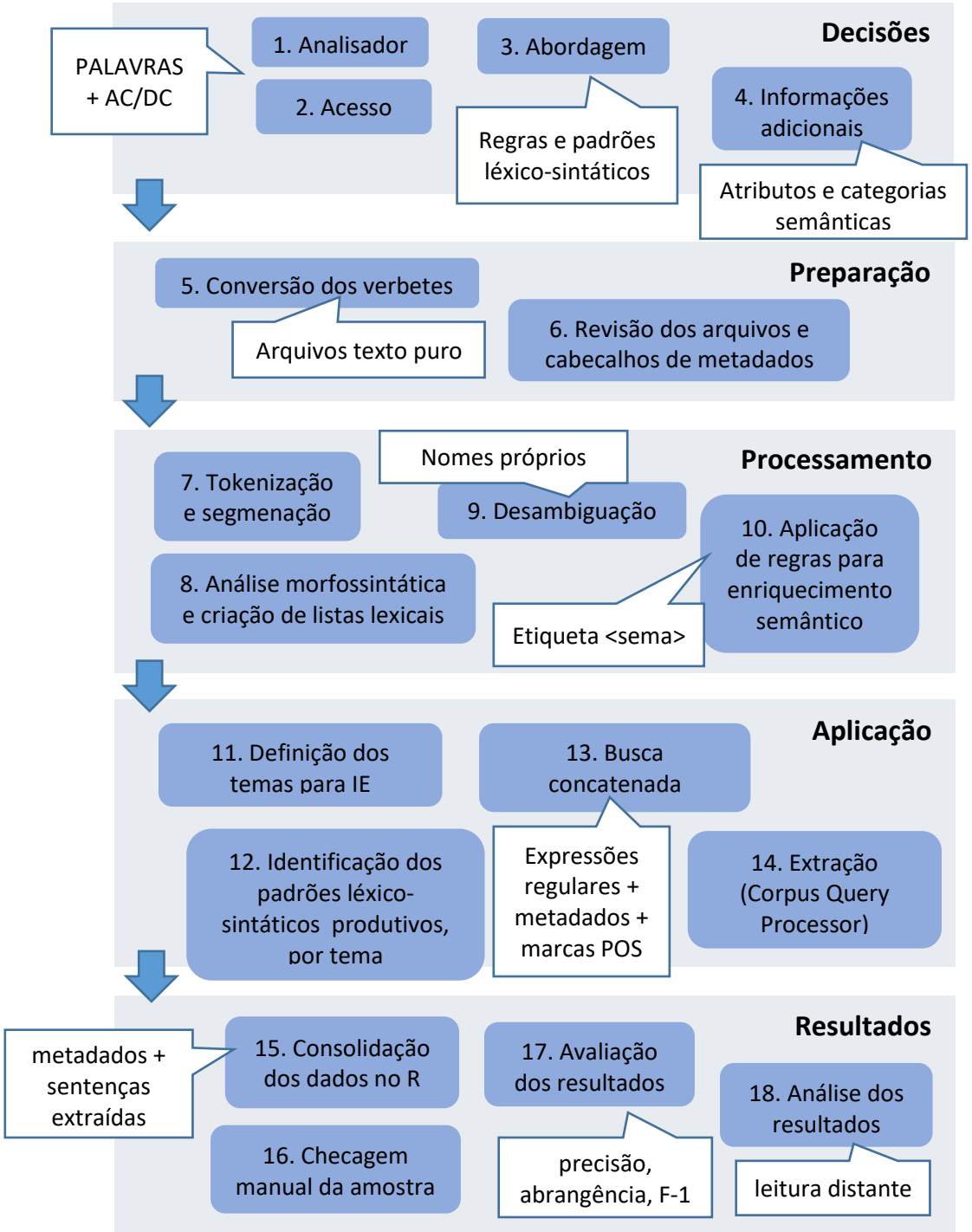

Figura 5 - Etapas do processo

5.1
Etapa 1: Decisões

Analisador: adoção do PALAVRAS

O parser utilizado para realizar a análise morfossintática do corpus DHBB é o PALAVRAS, escolhido por uma série de razões. As principais são que este analisador é especialmente voltado para a língua portuguesa, sendo considerado um dos melhores dentro da abordagem escolhida, com uma análise sintática e semântica de qualidade, e é também o adotado pela Linguateca para o processamento de todos os corpora incluídos no AC/DC[23] (Santos & Bick, 2000).

Segundo seu idealizador Eckard Bick (2007), o PALAVRAS é apoiado em regras, tanto localmente (no reconhecimento de padrões morfológicos) quanto globalmente (no contexto da sentença) baseando-se em um conjunto de ferramentas chamadas de *Constraint Grammars* (gramáticas de restrições), para a análise dos corpora.

A tarefa de NER é integrada à anotação gramatical, sendo que as marcações das entidades candidatas são feitas a partir de três níveis e desambiguadas através de regras CG: i) entradas lexicais conhecidas e listas de termos adicionais (com milhares de entradas); ii) predição baseada em padrões (módulo morfológico) e; iii) inferência baseada em contexto para palavras desconhecidas. Além disso, o parser une expressões fixas com a função sintática-semântica não composicional para MWEs, criando tokens compostos e facilitando para as regras sintáticas CG baseadas em tokens.

Nomes são tratados como MWEs e classes semânticas de NER são atribuídas ao todo, e não às partes, ajudando na tarefa de identificação das entidades, como nos casos “Rio de Janeiro”, “João Goulart”, “Dicionário Histórico-Biográfico Brasileiro”, “Comissão Parlamentar de Inquérito”, “Golpe de 1964”.

Ambiente de acesso: AC/DC

O projeto AC/DC é uma iniciativa da Linguateca para tornar corpora anotados acessíveis para o usuário comum através de uma interface web. O sistema

[23] Todas as etapas de processamento do DHBB para o AC/DC foram apoiadas pela profª Diana Santos, líder e fundadora do projeto Linguateca, a quem mais uma vez agradeço.

foi customizado a partir do IMS Open Corpus Workbench[24] (CWB), uma coleção de ferramentas *open source* voltada para interrogação de corpora enriquecidos de anotações linguísticas. O CWB vê o corpus como uma entidade com integridade própria, sobre o qual se pode interrogar, mas nunca alterar, permitindo diversas visões sobre o mesmo corpus ao fazer uso de vários níveis diferentes de anotação (Santos & Ranchhod, 1999).

Ao ser incluído no AC/DC, as consultas ao corpus DHBB conseguem retornar diferentes tipos de informação, todos eles relacionados com os conceitos de concordância, distribuição, distribuição cruzada, frequência e ordenação (ranking), e são feitas através da linguagem CQP (Corpus Query Processor)[25] que utiliza uma combinação de expressões regulares, propriedades linguísticas e atributos variáveis em sua sintaxe. É possível, por exemplo, pedir por verbos em determinados tempos ou pessoas, sintagmas com determinadas características, orações relativas onde o pronome relativo exerce uma determinada função sintática, ocorrência de adjetivos com determinados sufixos, dentre uma série de outras possibilidades (Freitas, Santos & Gonçalves, 2011). Além disso o sistema permite consulta a informação externa ao texto, no caso, aos metadados dos verbetes.

Para obter as sentenças resultantes da extração de informações, não acessamos a interface web, mas utilizamos diretamente o terminal CQP (Corpus Query Processor) da Linguateca. Nele, aplicamos as expressões de busca concatenadas e gravamos o resultado com todas as ocorrências em um arquivo texto.

Abordagem para extração automática de informação

O desenvolvimento deste trabalho parte de uma abordagem fortemente identificada pelos trabalhos desenvolvidos por Marti Hearst (1992, 1998), onde a informação é extraída a partir de um conjunto de padrões e regras léxico-sintáticos aplicados ao corpus, específicos ao domínio.

Apesar de sua simplicidade, a técnica apresenta resultados bastante precisos (Hearst, 1992; Makarov, 2018), com a vantagem de serem transparentes e

[24] http://cwb.sourceforge.net/
[25] Tutorial disponível em: http://cwb.sourceforge.net/files/CQP_Tutorial.pdf

prontamente examináveis, sendo acionados pela correspondência de estruturas específicas no texto. Por outro lado, a construção manual de padrões pode se mostrar de alto custo (Schrodt, 2006; Agt-Rickauer, 2019), pois não são facilmente transportáveis entre domínios, e qualquer adaptação a um novo domínio requer um grande esforço humano.

A escolha desta estratégia se deve, dentre outros fatores, à previsibilidade dos textos do DHBB, cuja escrita segue uma estrutura bastante padronizada. Além disso, o AC/DC permite que melhoremos a identificação de certos tipos de informação no corpus – de diferentes campos semânticos – através da criação de regras e léxicos específicos, conforme o interesse de pesquisa e de análise. Assim, progressivamente novas anotações são incorporadas ao corpus e acessadas normalmente nas expressões de busca.

Por último, sendo uma abordagem que não requer domínio computacional complexo na aplicação dos padrões – ao contrário dos métodos ancorados em *machine learning* –, mostra-se adequada para os fins exploratórios desta tese.

Informações adicionais: atributos e categorias semânticas

Para ajudar no cruzamento dos dados extraídos, criamos alguns atributos específicos para o DHBB (a maioria proveniente dos metadados dos verbetes) e os associamos a cada token do corpus. São eles:

- dicionário → o AC/DC abriga três dicionários distintos. Caso não seja especificado o dicionário, a busca varrerá todos eles, mas se o usuário desejar, pode selecionar apenas um: [dhbb] para o Dicionário histórico-biográfico brasileiro; [dhbpr] para o Dicionário histórico-biográfico da Primeira República; [dprrj] para o Dicionário da política republicana do Rio de Janeiro. Todos editados pelo CPDOC/FGV.
- classe → identifica se o verbete é [temático] ou [biográfico].
- fonte → indica o verbete em si que é exatamente o nome completo da entrada, como por exemplo: [José Ribamar Ferreira de Araújo Costa] ou [REVOLUÇÃO DE 1930]. Por agora os verbetes temáticos têm o nome em maiúsculas.
- entidade → identifica uma personalidade única, atribuída (por enquanto) apenas a nomes próprios humanos (todas as outras palavras têm id 0). Se essa personalidade

estiver descrita no DHBB, o valor da entidade é o número que corresponde ao id do verbete na base de origem, por exemplo: [dhbb_10], [dhbb_5458][26].

Se essa personalidade não estiver descrita no DHBB ou ainda não tiver sido identificada, o valor de id será "NS" (não sabemos) ou "NV" (não verbetado).

- sexo → [m] para biografias de homens, [f] para biografias de mulheres.
- cargos → identifica todos os principais cargos ou papéis associados ao biografado. Este atributo concatena em uma única string todos os valores encontrados no campo "cargos" oriundos do cabeçalho de metadados do verbete em questão, respeitando a sequência original. Assim, teremos para o verbete de Pedro Aleixo a seguinte string:

```
[const1934_depfedMG1935-1937_depfedMG1959-1966_minEduc1966_depfedMG1966-1967_vice-presRep1967-1969]
```

Que mostra que ele exerceu os cargos de constituinte, deputado federal por Minas Gerais em 3 mandatos não consecutivos, ministro da Educação e vice-presidente da República. Os períodos, quando houver, são discriminados junto aos cargos..

- autores → caso o verbete possua autoria, a informação é fornecida por este atributo.

Estes atributos foram adicionados a todos os tokens de um verbete biográfico. Isso significa que cada palavra/unidade terá agregado a si uma 'coluna' preenchida com o valor para dicionário, outra para fonte, outra para entidade, outra para cargos e assim por diante – o que nos permite, por exemplo, procurar quem "nasceu em Cataguazes" (utilizando uma busca léxico-sintática) e foi "presidente da República" (buscando pelo valor correspondente na coluna "cargos").

O PALAVRAS possui um conjunto pré-determinado de marcações semânticas utilizadas para identificar classes de nomes no corpus[27]. Os tipos que interessam para o contexto do DHBB são os seguintes:

- <hum> = pessoa
- <org> e <inst> = organização/instituição

[26] Cada verbete tem uma identificação única, e a lista completa encontra-se em: https://www.linguateca.pt/acesso/entidadesDHBB.html.
[27] A lista completa destas marcações pode ser consultada na página: https://visl.sdu.dk/visl/pt/info/portsymbol.html

- <party> = partido político
- <occ> e <event> = evento/acontecimento
- <civ> = lugar
- <tit> = documento/obra
- <hprof> = cargo/papel

No entanto, identificamos uma classe semântica bastante relevante para o domínio, relacionada a planos de governo, programas, acordos, tratados, leis, decretos, códigos etc., porém que não existe no analisador. Assim, criamos uma classe específica para este fim:

- <titfpol> = formulações políticas

5.2 Etapa 2: Preparação

A etapa de preparação do corpus visa tratar os textos de maneira a torná-los passíveis de serem lidos pelo analisador. No caso do DHBB, todo o conteúdo provém de um sistema de informações integrado aos outros acervos que a instituição curadora abriga. A estrutura do seu banco de dados é constituída de um campo de texto para acolher os verbetes codificados em html e alguns poucos metadados, basicamente título e tipo (temático ou biográfico). O processo e a lógica adotada para a extração destes dados da base visando o PLN são descritos por (Paiva et al, 2014) e (Rademaker et al, 2015), no âmbito de um projeto maior de reestruturação dos sistemas de informação do acervo do CPDOC.

Arquivos processáveis e cabeçalhos de metadados

O primeiro passo foi converter cada entrada em um arquivo texto no formato markdown, uma sintaxe de marcação leve e legível por máquina (Gruber, 2004). Foi necessário realizar vários procedimentos de limpeza para excluir marcações de html e de controle de formatação gerados pelo editor eletrônico utilizado na redação dos verbetes. A decisão de adotar o formato de texto puro considera uma série de vantagens, tais como facilidade na manutenção dos arquivos, independência de plataforma, formato ideal para PLN e agilidade no enriquecimento de metadados (Rademaker et al, 2015),

Assim, cada verbete se tornou um arquivo identificado por um número sequencial – originalmente, o id da base – e ganhou uma área em seu cabeçalho para a anotação estrutural. A revisão manual de cada entrada permitiu a normalização desta área com os seguintes metadados:

```
---
title: GOULART, João
natureza: biográfico
sexo: m
cargos:
 - dep. fed. RS 1951
 - dep. fed. RS 1952-1953
 - min. Trab. 1953-1954
 - dep. fed. RS 1954
 - vice-pres. Rep. 1956-1961
 - pres. Rep. 1961-1964
autor:
 - Marieta de Morais Ferreira
---
```

Tabela 7 - Exemplo de cabeçalho de um arquivo de verbete

As primeiras linhas informam o nome do verbete, a natureza, e, caso seja do tipo biográfico, o gênero e os cargos que o personagem exerceu, dispostos de forma abreviada e em ordem cronológica. No caso acima, João Goulart foi deputado federal em 1951, novamente de 1952 a 1953, depois ministro do Trabalho entre 1953 e 1954, de novo deputado federal em 1954, vice-presidente da República entre 1956 a 1961, e finalmente presidente da República, entre 1961 e 1964.

Conforme veremos na etapa de aplicação, os metadados têm papel fundamental na extração de informação do DHBB, promovendo cruzamentos de dados interessantes. Por outro lado, mesmo sabendo que as fontes de referência são um dos elementos importantes de qualquer produção científica, servindo neste caso como garantia literária para os verbetes, elas não foram incluídas no corpus. A grande quantidade de nomes próprios encontrados irrelevantes para a pesquisa impactaria sobremaneira na tarefa de NER e por consequência, na extração de informação envolvendo essas entidades.

Todos os arquivos dos verbetes foram disponibilizados no github[28].

[28] https://github.com/cpdoc/dhbb

5.3
Etapa 3: Processamento

Nesta seção são descritos os processos percorridos para a construção do corpus anotado, com uma síntese da lógica e recursos utilizados.

5.3.1
Tokenização e segmentação

O processo de tokenização (ou atomização) do DHBB seguiu o mesmo pipeline adotado por todos os outros corpora incorporados ao AC/DC[29], prevendo a identificação das palavras básicas, dos sinais de pontuação, e a marcação de frases e parágrafos.

De forma bem resumida, contrações, palavras compostas, palavras hifenizadas, verbos com clíticos, abreviaturas com sinais de pontuação e os próprios sinais de pontuação são considerados como sendo tokens separados, e para cada uma destas definições há regras visando identificá-los. A separação das frases também segue uma série de tratamentos capazes de reconhecer, por exemplo, quando a presença de uma maiúscula não significa início de uma frase, quando um ponto não é ponto final e sim parte de uma palavra, quando uma frase é citação dentro de outra frase etc.

Na versão 7.3 do DHBB[30], temos os seguintes números:

Frases	457.101
Tokens	15.811.411
Palavras	13.910.760
MWEs	155.620

Tabela 8 - Dimensão do corpus DHBB

Lembrando que o Dicionário é atualizado periodicamente e estes totais podem variar conforme a versão disponibilizada online.

[29] Descrito nesta página: https://www.linguateca.pt/acesso/atomizacao.html
[30] https://www.linguateca.pt/acesso/contabilizacao.php#dhbb

5.3.2
Análise morfossintática e anotação

Nesta etapa uma versão anotada do DHBB é criada a partir da análise pelo PALAVRAS e de outros processamentos comuns do projeto AC/DC. A estruturação de cada verbete – agora em um formato pseudo-xml – além de fazer a separação de parágrafos e frases, incorpora metadados aos verbetes e atribui marcas sintáticas e semânticas aos tokens. Também é nesta fase que os arquivos são agregados para serem processados como um corpus único[31].

Abaixo temos um exemplo de saída do PALAVRAS para o excerto "no entanto, o PT logrou êxito e a candidata Dilma Rousseff foi eleita presidente", extraído do verbete Fernando da Mata Pimentel[32]:

Forma	Lema	Marca interna	POS	F. sintática
no entanto	[no=entanto]	<kc>	ADV	@ADVL>
,				
O	[o]	<artd>	DET M S	@>N
PT	[PT]	<party> <*>	PROP M S	@SUBJ>
Logrou	[lograr]	<fmc>	V PS 3S IND VFIN	@FMV
êxito	[êxito]	<act>	N M S	@<ACC
E	[e]	<co-vfin> <co-fmc>	KC	@CO
A	[o]	<artd>	DET F S	@>N
candidata	[candidato]	<Hprof>	N F S	@SUBJ>
Dilma Rousseff	[Dilma=Rousseff]	<hum> <*>	PROP F S	@N<
foi	[ser]	<fmc>	V PS 3S IND VFIN	@FAUX
Eleita	[eleger]	<vt>	V PCP F S	@IMV @#ICL-AUX<
presidente	[presidente]	<Hprof>	N M/F S	@<SC

Tabela 9 - Marcações morfossintáticas do PALAVRAS a uma frase em português

[31] Para uma descrição mais detalhada deste processo de estruturação do material, o artigo de Santos & Bick (2000) pode ser consultado.
[32] Acessível em: http://www.fgv.br/cpdoc/acervo/dicionarios/verbete-biografico/pimentel-fernando

A cada unidade é atribuído o seu lema, a sua categoria gramatical e outros atributos morfossintáticos e a sua função sintática[33].

Na primeira coluna está o texto enviado, na segunda, a forma canônica (lema) e nas seguintes as marcações morfossintáticas. Note que o PALAVRAS consegue identificar locuções ("no entanto") e nomes próprios ("Dilma Rousseff") e as trata como unidades. O lexema "logrou" tem sua forma canônica "lograr" e classificação V PS 3S IND VFIN, ou seja, trata-se de um verbo, no pretérito perfeito simples, 3ª pessoa singular, modo indicativo, na forma finita. Além disso, a marca @FMV indica a atribuição da função sintática como sendo ele o núcleo do predicado verbal (*finite main verb*).

O formato AC/DC, que é o formato final após o corpus ter sido processado e tratado, resulta em um conjunto de campos separados por caracteres de tabulação. O quadro abaixo mostra como ficou o mesmo texto "no entanto, o PT logrou êxito e a candidata Dilma Rousseff foi eleita presidente."

<mwe lema=no=entanto pos=ADV>	
no	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 em+o PRP+DET_artd 0 S M ADVL>+>N
entanto	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 entanto ADV 0 0 0 ADVL>
</mwe>	
,	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 , PU 0 0 0 PONT
o	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 o DET_artd 0 S M >N
PT	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 PT PROP_party 0 S M SUBJ>
Logrou	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 lograr V PS_IND 3S 0 FS-STA
Êxito	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 êxito N 0 S M <ACC

[33] Para uma visão geral de todas as etiquetas que o PALAVRAS atribui ao corpus, consultar: http://visl.sdu.dk/visl/pt/info/

e	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 e KC 0 0 0 CO
a	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 o DET_artd 0 S F >N
candidata	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 candidata N 0 S F SUBJ>
Dilma	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 11455 Dilma=Rousseff PROP_hum 0 S F N<
Rousseff	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 11455 Dilma=Rousseff PROP_hum 0 S F N<
foi	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 ser V-AUX_aux_fmc_cjt PS_IND 3S 0 FS-STA
eleita	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 eleger V PS_PASSIVA_COMP_IND 3S F ICL-AUX<
presidente	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 presidente N 0 S M/F <SC
.	Fernando Damata Pimentel prefBeloHorizonte2003-2009_minDesenvIndCom2011-2014 . PU 0 0 0 PONT

Tabela 10 - Exemplo de saída do formato AC/DC

Apenas os metadados nome do verbete (fonte) e cargos ocupados estão sendo mostrados na anotação; as outras foram omitidas por questão de simplificação. Mas é possível perceber como elas são incluídas, ficando agregadas a cada token individualmente, junto com os outros atributos de POS.

O resultado é um corpus contendo a identificação das palavras (*word*), lemas (*lema*), categorias gramaticais (*pos*), tempos verbais (*temcagr*), pessoa e número (*pessnum*), função sintática (*func*) e informações semânticas adicionais (*sema*).

5.3.3
Criação de listas lexicais de entidades

Como é de se esperar, analisadores de um modo geral, não conseguem atribuir de forma 100% correta a classe semântica dos nomes próprios presentes em

um corpus. Para melhorar a identificação, criamos listas com entidades revisadas manualmente.

Três abordagens foram adotadas para a obtenção destas listas: i) uso de padrões ou pistas lexicais para buscar instâncias no corpus, ii) índices existentes na Wikipédia e iii) ocorrências identificadas pelo PALAVRAS e revistas manualmente.

Na primeira abordagem, aplicamos técnicas de extração consistindo na utilização de padrões lexicais. Assim, a construção "*presidiu* [o|a] [termo iniciando em letra maiúscula]" identifica instâncias de empresas e instituições; palavras como "*Colégio*", "*Universidade*" e "*Associação*" recuperam instâncias do tipo "*Colégio Pedro II*", "*Universidade Federal Fluminense*", "*Associação Brasileira de Empresas de Rádio e Televisão*"; expressões combinadas como "*localizado em* [termo em letra maiúscula]" e "*nasceu em* [termo em letra maiúscula]", tem grandes chances de recuperar instâncias do tipo lugar.

Para esta tarefa de mineração utilizamos o AntConc[34], ferramenta gratuita para análise de texto e geração de linhas de concordância. Com algumas poucas construções lexicais básicas recuperamos cerca de 7 mil instâncias do tipo organização, documento e profissão.

No caso das entidades do tipo evento percebemos que a estratégia com o AntCon não foi muito boa devido à falta de padrão das construções em que suas instâncias costumam aparecer. Neste caso, recorremos à Wikipédia, tirando proveito de que o domínio em questão – História do Brasil – é um campo tipicamente enciclopédico. Para tanto, não procuramos por verbetes específicos, mas fomos diretamente nas páginas do tipo "categorias", que funcionam como índices. A partir dos eventos listados nas categorias "Revoltas no Brasil", "Movimentos emancipacionistas do Brasil", "Movimentos separatistas no Brasil", "Rebeliões prisionais no Brasil", "Movimentos Emancipacionistas", "Lutas e Revoluções no Brasil", "Manifestações e protestos no Brasil" e "Movimentos do Brasil" obtivemos cerca de 230 instâncias de eventos.

[34] https://www.laurenceanthony.net/software/antconc/

Continuando, a terceira abordagem aproveita-se da anotação que o próprio PALAVRAS faz para identificar os nomes próprios. Por meio da interface do AC/DC, foram realizadas consultas linguísticas capazes de indicar a presença de nomes do tipo organização <org>, o que resultou em uma lista inicial de quase 28 mil instâncias. Manualmente revisamos cada entrada para validar a atribuição semântica recebida, e após esse processo o que restou foi um conjunto de cerca de 22.500 nomes, que agora podemos considerar como sendo uma lista “golden" de organizações. Nomes identificados erroneamente alimentaram outras listas: pessoa, formulação política, evento, lugar, documento, papel, partido (que consideramos um subtipo de organização).

- Organizações (25.970) → [sema=”org”]
- Formulações políticas (1.250) → [sema="titfpol"]
- Pessoas (18.488) → [sema="hum"]
- Lugares (140) → [sema=”civ”]
- Eventos (350) → [sema=”evento”]
- Partidos (1.011) → [sema=”party”]

Estas são as listas “golden" que se tornaram um atributo [sema] no AC/DC e que pretendemos ir melhorando de forma contínua. É preciso reforçar que a lista de organizações contempla muito mais instâncias que as outras porque foi o tipo de entidade que extraímos do corpus e checamos manualmente neste primeiro exercício de criação de léxicos.

5.3.4
Problemas de segmentação de nomes próprios

Muitas vezes, ao checar de perto as informações recuperadas no corpus, nos deparamos com alguns erros de segmentação que o analisador não foi capaz de identificar. Um exemplo é o caso de “*Eugênia Lopes de Oliveira Prestes de Macedo Soares*”, que foi reconhecido automaticamente como sendo dois nomes próprios ao invés de um: “*Eugênia Lopes de Oliveira Prestes*” e “*Macedo Soares*”. Em outro exemplo, “Simpatizante de Vargas” deveria reconhecer apenas “Vargas”, mas teve “Simpatizante” agregado ao seu nome.

Erros assim são corrigidos por uma ferramenta chamada “Corte-e-costura”. O “Corte-e-costura” baseia-se em regras e foi criado no âmbito do AC/DC com o objetivo inicial de ajudar na anotação de corpus com atributos semânticos específicas (Mota, 2014)[35]. Mas outras funcionalidades foram sendo acrescentadas, aumentando-se a expressividade das regras. Assim, as regras passaram a ser usadas para juntar expressões ou entidades com mais de uma palavra, correspondendo, do ponto de vista do formato AC/DC, à marcação do atributo estrutural <mwe>.

O formato das regras é bastante similar à sintaxe do CQP utilizada no AC/DC, o que torna mais fácil para os linguistas que já têm experiência com o sistema fazer a revisão do seu corpus, de forma iterativa. Abaixo temos dois exemplos de regra:

```
# (1) Regra para dizer que "Golpe de 64" é uma MWE:

a:[word="Golpe"] b:[word="de"] c:[word="64|1964"] >> <mwe
pos="PROP_occ"> a: b: c: </mwe>

# (2) Regra para corrigir o lema Simpatizante=de=Vargas:

a:[lema="Simpatizante=de=Vargas"]
b:[lema="Simpatizante=de=Vargas"]
c:[lema="Simpatizante=de=Vargas"] >> a:[lema="simpatizante"
& pos="N"] b:[word="de" & pos="PRP"] c:[lema="Vargas"]
```

Regras para modificar atributos de palavras e lemas sempre apresentam um antecedente e um consequente, delimitados pelos caracteres “>>”. Na regra (1) acima estamos juntando as palavras “Golpe”, “de” e “64” ou “1964” em um único lema, e atribuindo a este a classe de nome próprio do tipo evento. Na regra (2), queremos desacoplar a palavra “simpatizante” do nome próprio “Vargas”, então o que ela faz é: ao encontrar a sequência “Simpatizante de Vargas” no corpus, o programa adiciona – ou substitui, caso já exista o atributo – a classe gramatical

[35] Vários experimentos nesse sentido foram realizados, por exemplo, sobre partes do corpo humano (Freitas, 2013), sobre emoções (Mota, 2014), sobre vestuário (Santos et al, 2009) e sobre o campo semântico de cor (Silva e Santos, 2009).

substantivo ao primeiro termo (pos="N"), a classe gramatical preposição ao segundo (pos="PRP") e mantém como nome próprio o terceiro, que é Vargas. Neste último exemplo, como estamos trabalhando com um lema constituído originalmente por três termos, a sentença será recuperada três vezes nas regras[36].

5.3.5
Desambiguação e correspondência entre entidades

O trabalho de desambiguação é conhecido em inglês como *entity grounding*[37] ou *entity disambiguation*[38], e através dele criamos associações entre as formas variadas de escrita de uma determinada entidade (lemas), atribuindo a elas uma identificação única. No DHBB, a tarefa foi identificar e fazer a correspondência apenas entre os diferentes nomes próprios das pessoas biografadas. O processo aconteceu da forma descrita a seguir.

Por via de regra, o nome completo da pessoa biografada virá sempre na primeira linha do texto do verbete delimitado pelos caracteres « e ».

```
«Álvaro Augusto Ribeiro da Costa» nasceu em Fortaleza em 17 de abril de 1947.
```

Será este nome que utilizaremos como nosso identificador único, ou seja, o referente a que corresponderão todas as possíveis derivações do nome da pessoa em questão, sem esquecer que ele também está associado ao código ID do próprio verbete.

Para saber quais nomes precisam ser correspondidos, usamos a expressão de busca abaixo, pedindo por distribuição por lemas:

```
[dicionario="dhbb" & pos=".*PROP" & sema=".*hum.*" & entidade!="dhbb_.*"]
```

Dessa forma recuperamos todos os nomes próprios a que o PALAVRAS atribuiu como sendo "humano", e que não foram identificados como um titular de verbete (ou seja, não existe como uma entidade).

[36] O arquivo de regras para corrigir as segmentações erradas no DHBB encontra-se em: https://www.linguateca.pt/acesso/corpos/dhbb/regras_corr_PALAVRAS_roupa.excl
[37] https://en.wikipedia.org/wiki/Symbol_grounding_problem
[38] https://en.wikipedia.org/wiki/Entity_linking

Distribuição

Houve **33633** valores diferentes de **lema**. Apresentam-se apenas 999, por ordem decrescente de frequência

Getúlio	987
Sarney	932
Salvador	792
Washington	774
Paris	486
França	437
Costa	432
Santa=Maria	366
Nilo=Peçanha	356
Bernardes	333
Santos	320
Juarez	307
Campos	291
João=Paulo	248
Pedro=Collor	236
Consulta=dos=Ministros=das=Relações=Exteriores	234
Vitória	227
Itamar	219
Lisboa	219
Hélio=Silva	214
Segurança=Pública	208
Álvaro=Dias	188

Figura 6 - Resultado da busca por lemas sem correspondência

Da lista resultante, é feita uma checagem manual indo aos trechos onde estes lemas são mencionados, a fim de evitar correspondências erradas. Alguns nomes são óbvios, como por ex. “Lula” (que sempre irá se referir a “Luís Inácio da Silva”) e “Getúlio=Vargas” (sempre será “Getúlio Dornelles Vargas”). Mas alguns, como “Vargas” e “Castelo=Branco”, podem se referir a mais de uma pessoa.

AC/DC lema (lema)	Nome completo de uma entrada no DHBB (entidade)
Aécio=Neves	Aécio Neves da Cunha
Alencar=Castelo=Branco	Humberto de Alencar Castelo Branco
Anthony=Garotinho	Anthony William Matheus de Oliveira
Getúlio=Vargas	Getúlio Dornelles Vargas
Lula	Luis Inácio da Silva

Tabela 11 – *Grounding* - correspondência entre entidades

Na tabela acima, temos na primeira coluna os lemas extraídos do corpus que precisam ser associados às entidades da segunda coluna.

Após a checagem, uma tabela de correspondência é criada indicando as correspondências entre os lemas e os ids dos verbetes. Em geral, quando a

ocorrência se verifica em um verbete próprio (por ex., “Vargas” sendo mencionado no verbete de “José Israel ‘Vargas”), se está a referir ao seu titular (“José Israel Vargas”). Mas não podemos tomar esta regra como certa, pois existem casos em que os lemas podem se referir a outras pessoas. No exemplo abaixo, os dois primeiros seguem a regra (o lema se refere ao próprio titular do verbete), mas os seguintes não:

AC/DC lema (X)	Verbete onde ocorre (Z)	Verbete (entidade) a que corresponde (Y)
Vargas	Getúlio Dornelles Vargas	Getúlio Dornelles Vargas
Vargas	Jose Israel Vargas	Jose Israel Vargas
Vargas	Alzira Vargas do Amaral Peixoto	Getúlio Dornelles Vargas
Vargas	Benjamim Dornelles Vargas	Getúlio Dornelles Vargas
Vargas	Lutero Sarmanho Vargas	Getúlio Dornelles Vargas

Tabela 11 - O verbete de Alzira Vargas do Amaral Peixoto menciona Vargas para se referir ao presidente Getúlio Dornelles Vargas, que também é seu pai

Assim, após uma verificação manual destas menções, implementamos uma forma específica de regras de correspondência que inclui exceções, conforme exibido na tabela acima. As regras devem ser lidas como “designação X deve ser correspondido à entidade Y se ela aparecer na entrada Z". Através de um script, as regras são aplicadas no corpus, utilizando a mesma metodologia baseada no “Corte e costura”.

Este é um processo iterativo: obtém-se a lista de lemas, fazem-se as correspondências; obtém-se nova lista, novas correspondências e assim sucessivamente. Sempre que se cria o corpus de novo, o processo de construção passa por um filtro que adiciona os identificadores, com base na lista de correspondências.

A tabela abaixo mostra como o *grounding* impactou no total de menções aos biografados no DHBB.

Processo iterativo	Total de lemas identificados como nomes de pessoas biografadas
Antes da aplicação das regras de correspondência	89.937
Primeira lista com 116 correspondências	147.085
Segunda lista com 71 correspondências	166.059

Tabela 12 - Resultados iniciais do processo de grounding

Antes de aplicar as regras de correspondência, o número de ocorrências era de 89.937. Com a produção de uma primeira lista contendo 116 regras de correspondência na forma ilustrada na tabela 11, conseguimos aumentar para 147.085. Em uma segunda iteração, adicionando 71 novas correspondências, obtivemos 166.059 casos.

5.3.6 Enriquecimento semântico do corpus

O enriquecimento semântico é feito a partir da anotação semântica no corpus, um processo semiautomático (Santos & Mota, 2010), utilizando uma ferramenta e metodologia desenvolvidas para este fim. Vários experimentos nesse sentido foram realizados no AC/DC, como por exemplo, sobre partes do corpo humano (Freitas, 2013), sobre emoções (Mota, 2014), sobre vestuário (Santos et al, 2009) e sobre o campo semântico de cor (Silva e Santos, 2009).

Em termos gerais, o processo parte de um léxico inicial que é aplicado às palavras do corpus, anotando-as como pertencentes ao campo semântico em questão. Em seguida, por meio da análise de contexto das palavras anotadas, são criadas regras de especialização ou de eliminação, para corrigir casos ambíguos.

A motivação para a anotação costuma vir de questões colocadas à priori, sobre o que se deseja extrair de informação do corpus. No nosso caso, vem do interesse em identificar quem são os políticos que detêm vínculos familiares com outros políticos e saber que vínculos seriam esses. A seguir, descrevemos a metodologia aplicada no DHBB.

Anotação semântica sobre relações familiares

Em primeiro lugar definimos um léxico com palavras pertencentes ao domínio de parentesco. Separamos as palavras em três classes: laços, dinâmica e

especificada. A primeira abrange termos que descrevem laços de família, quer legais quer sociais, quer de sangue quer por afinidade; a segunda, termos que remetam ao processo de constituição ou dissolução de um laço, ou ainda a uma consequência ou ausência dele; já a terceira classe reúne termos que em geral se apresentam associados aos primeiros, como um qualificador:

(1) [familia:lacos]:

afilhada, afilhado, amante, amásia, antepassado, ascendente, avó, avô, bisavó, bisavô, bisneta, bisneto, casal, comadre, compadre, companheira, companheiro, concubina, cônjuge, cunhada, cunhado, descendente, enteada, enteado, esposa, esposo, ex-esposa, ex-esposo, ex-marido, ex-mulher, ex-namorada, ex-namorado, familiar, família, filha, filho, genro, herdeira, herdeiro, irmã, irmão, madrasta, madrinha, mamã, mamãe, mana, mano, marido, mãe, meia-irmã, meio-irmão, mulher, neta, neto, noiva, noivo, nora, padrasto, padrinho, pai, papai, papá, parenta, parente, prima, primo, prole, progenitor, progenitora, sobrinha, sobrinha-neta, sobrinho, sobrinho-neto, sogra, sogro, tataravó, tataravô, tetravó, tetravô, tia, tia-avó, tio, tio-avô, titi, titia, titio, trisavó, trisavô,

(2) [familia:dinâmica]:

adoptar, adotar, amancebar, amasiar, amigar, apadrinhar, aparentar, casamento, casar, contrair [matrimônio, núpcias, casamento], desposar, desquitar, divorciar, enviuvar, matrimónio, matrimônio, noivado, noivar, núpcia(s), órfã, órfão, perfilhar, reconhecer, separar, solteiro, solteirice, viuvez

(3) [familia:especificada]:

adoptivo, adotivo, afastado, bastarda, bastardo, biológico, caçula, carnal, colaço, consanguíneo, consorte, de criação, de sangue, distante, em N grau, gémea, gêmea, gémeo, gêmeo, legítimo, ilegítimo, longe, materno, meio, morgada, morgado, natural, paterno, perto, por afinidade, por consideração, primogénito,

Assim, os termos em [família: especificada] costumam se apresentar associados a termos de laços de família, mas podem aparecer sozinhos também (e

nesse caso tornam-se familia:laços), como [filho] *caçula* ou [filha] *bastarda*, ajudando na detecção de expressões comuns ao domínio.

Essas informações seguem para um arquivo de regras[39] em formato VISLCG3 e são aplicadas ao corpus:

```
DELIMITERS="<.>" "<!>" "<?>" "<Â>" "<$!>" "<$.>" "<$?>";

MAPPING-PREFIX = %;

SETS

# Palavras que denotam relações de parentesco:

LIST LACOS = "afilhada", "afilhado", "amante", "amásia",
"antepassado", "ascendente", "avó", "avô", "bisavó", "bisavô",
"bisneta", "bisneto", "casal", "comadre", "compadre", "companheira",
"companheiro", "concubina", "cônjuge", "cunhada", "cunhado",
"descendente", "enteada", "enteado", "esposa", "esposo", "ex-esposa",
"ex-esposo", "ex-marido", "ex-mulher", "ex-namorada", "ex-namorado",
"familiar", "família", "filha", "filho", "genro", "herdeira",
"herdeiro", "irmã", "irmão", "madrasta", "madrinha", "mamã", "mamãe",
"mana", "mano", "marido", "mãe", "meia-irmã", "meio-irmão", "mulher",
"neta", "neto", "noiva", "noivo", "nora", "padrasto", "padrinho",
"pai", "papai", "papá", "parenta", "parente", "prima", "primo",
"prole", "progenitor", "progenitora", "sobrinha", "sobrinha-neta",
"sobrinho", "sobrinho-neto", "sogra", "sogro", "tataravó",
"tataravô", "tetravó", "tetravô", "tia", "tia-avó", "tio", "tio-avô",
"titi", "titia", "titio", "trisavó", "trisavô", "trisneta",
"trisneto", "viúva", "viúvo", "vovó", "vovô";
```

Tabela 13 - Trecho do arquivo de regras VISLCG3 para anotação semântica

Em alguns casos é preciso especificar os contextos gerais de uso, como no caso de “mulher” e “separado”, termos bastante usados fora do contexto de família (por exemplo, “primeira *mulher* a ser eleita” e “teve encontro *separado* com Roosevelt”). Para esta desambiguação também utilizamos regras.

[39] https://www.linguateca.pt/acesso/corpos/vislcg3/regras_emodizer.utf8.txt

```
# adição das palavras relacionadas com o campo da família: parentesco
etc.

ADD (%FAMLACOS) TARGET FAMILIA:LACOS;
ADD (%FAMESPECIFICADA) TARGET FAMILIA:ESPECIFICADA (0 (ADJ));
ADD (%FAMDINAMICA) TARGET FAMILIA:DINAMICA;
# tratamento da palavra "mulher"
ADD (%FAMLACOS) TARGET MULHERESPOSA (1 ("de")) (2 (PROP));
ADD (%FAMLACOS) TARGET MULHERESPOSA (-1 ("meu"));
ADD (%FAMLACOS) TARGET MULHERESPOSA (-1 ("seu"));
ADD (%FAMLACOS) TARGET MULHERESPOSA (-1 ("teu"));
# tratamento da palavra "ascendente"
ADD (%FAMLACOS) TARGET LACOSN (0 (N));
# tratamento do verbo "separar"
ADD (%FAMDINAMICA) TARGET SEPARAR (0 (PS)) (1 ("de")) (2 (PROP));
ADD (%FAMDINAMICA) TARGET SEPARAR (0 (PS)) (1 ("se")) (2 ("de"));
ADD (%FAMDINAMICA) TARGET SEPARAR (0 (PCP)) (1 ("de")) (2 (PROP));
# tratamento do verbo "aparentar"
ADD (%FAMDINAMICA) TARGET APARENTAR (0 (PCP));
# tratamento do verbo "contrair"
ADD (%FAMDINAMICA) TARGET CONTRAIR (1 ("matrimônio"));
ADD (%FAMDINAMICA) TARGET CONTRAIR (1 ("matrimónio"));
ADD (%FAMDINAMICA) TARGET CONTRAIR (1 ("casamento"));
# tratamento das palavras de especificação que estão sozinhas
ADD (%FAMLACOS) TARGET FAMILIA:ESPECIFICADA (0 (N));
```

Tabela 14 - Trecho do arquivo de regras VISLCG3 para desambiguação de palavras para anotação semântica

Feito isso, as regras são aplicadas ao corpus e o atributo <sema> pode ser utilizado para a identificação de estruturas contendo relações familiares entre políticos do DHBB:

```
[dicionario="dhbb" & entidade="dhbb.*"]+ [: entidade!="dhbb.*" :] ","
[sema="familia:lacos.*"] [pos="PRP.*"] [entidade="dhbb.*"]+ [: entidade!="dhbb.*" :]
```

4909 : O controle pefelista do Congresso se completaria com a eleição de **Luís Eduardo Magalhães, filho de Antônio Carlos**, para a presidência da Câmara .

4910 : Este grupo tinha como principais membros os deputados federais filiados ao PFL maranhense: Sarney Filho, César Bandeira, Costa Ferreira, José Reinaldo Tavares e **Ricardo Murad, cunhado de Roseana**, e ainda Paulo Marinho, do Partido Social Cristão (PSC) , e Nan Sousa, do Partido Social Trabalhista (PST) .

4915 : Representante no setor econômico do Partido do Movimento Democrático Brasileiro (PMDB) , base de sustentação do governo, passou a dividir decisões com o dministro da Fazenda, **Francisco Dornelles, sobrinho de Tancredo** e representante do Partido da Frente Liberal (PFL) que, com a nova situação, perdeu força política .

4922 : No Rio Grande do Sul, o governador **Leonel Brizola, cunhado de João Goulart** e fiel à Constituição, organizou um movimento de resistência à oposição militar lançando a « campanha da legalidade », com o objetivo de assegurar a posse do vice-presidente .

5067 : Seu sobrinho, **Jorge Roberto Silveira, filho de Roberto Silveira**, foi deputado estadual no Rio de Janeiro entre 1979 e 1987, e prefeito de Niterói entre 1989 e 1993, conquistando novo mandato para o período 1997-2001 .

Figura 7 - Identificação de relações familiares entre políticos que possuem entrada no DHBB

O fragmento acima mostra em contexto alguns dos diversos casos de relações familiares existentes entre os políticos. Vale lembrar que foram aplicadas regras de correspondência entre os nomes dos biografados para que eles pudessem ser recuperados sem ambiguidades (por ex., Antônio Carlos, no caso acima, se refere ao político Antônio Carlos Peixoto de Magalhães).

Este processo é explicado na seção 5.3.5, sobre desambiguação e correspondência entre entidades. No total são 6.351 verbetes biográficos contendo uma ou mais menções a termos de família. Somadas, as menções chegam a 20.788, distribuídas por 80 diferentes tipos, cujas primeiras vinte posições seguem na tabela abaixo:

Laço familiar	N. menções	Laço familiar	N. menções
filho	10.808	sobrinho	214
pai	1.739	parente	185
irmão	1.521	casal	152
família	1.486	irmã	149

filha	1.330	marido	147
companheiro	431	avô	134
tio	364	cunhado	122
primo	327	familiar	109
esposa	289	sogro	106
mãe	264	mulher	101

Tabela 15 - As vinte relações familiares mais citadas nos verbetes biográficos do DHBB

Previsivelmente a relação "filho" é, de longe, a que aparece em primeiro lugar, pois por via de regra, filiação de pai e mãe é uma informação que deve constar em toda biografia. Outras, como "família" e "companheiro" devem ser olhadas com cuidado sob o risco de estarem em contextos fora de laços familiares: "o número de *famílias* assentadas pelo governo", "seu *companheiro* de chapa". Veremos casos assim mais adiante.

Um dos recursos disponíveis no AC/DC é o Distribuidor[40], uma interface que permite obter distribuições de frequências conjuntas. Para saber quantas e quais relações comparecem em cada verbete biográfico, utilizamos a expressão de busca:

```
?dicionario=/dhbb/ ?classe=/biográfico/ ?sema=/familia:lacos.*/ fonte lema
```

O resultado é a lista de frequência em números absolutos e percentuais, como no exemplo abaixo.

[40] https://www.linguateca.pt/

fonte	Frequência Total	lema	Frequência Parcial	%
Getúlio Dornelles Vargas	106	compadre	1	0,94%
		companheiro	11	10,38%
		enteada	1	0,94%
		esposa	1	0,94%
		familiar	5	4,72%
		família	11	10,38%
		filha	4	3,77%
		filho	20	18,87%
		genro	2	1,89%
		herdeiro	2	1,89%
		irmão	21	19,81%
		marido	2	1,89%
		mulher	1	0,94%
		mãe	1	0,94%
		neta	1	0,94%
		pai	11	10,38%
		parente	3	2,83%
		primo	2	1,89%
		sobrinha	2	1,89%
		sobrinho	1	0,94%
		tio	3	2,83%

Tabela 16 - distribuição das ocorrências de relações familiares no verbete de Getúlio Vargas

Getúlio Dornelles Vargas lidera o ranking de relações com 106 menções distribuídas por 21 diferentes tipos, sendo “irmão” e “filho” os primeiros da lista. O quadro a seguir é uma compilação dos 20 primeiros verbetes no ranking de frequência:

Verbete	Frequência total	Tipos de relação	Verbete	Frequência total	Tipos de relação
Getúlio Dornelles Vargas	106	21	Raul Belens Jungmann Pinto	35	6
Luís Carlos Prestes	54	18	Agildo da Gama Barata Ribeiro	34	11
Osvaldo Euclides de Sousa Aranha	53	19	João Belchior Marques Goulart	34	14
José Ribamar Ferreira de Araújo Costa	53	20	Antônio Carlos Ribeiro de Andrada	33	15
Luís Inácio da Silva	41	9	Leonel de Moura Brizola	33	13

Eurico Gaspar Dutra	40	17	Afonso Arinos de Melo Franco	33	8
Fernando Afonso Collor de Melo	39	12	Pedro Aurélio de Góis Monteiro	31	11
Alzira Vargas do Amaral Peixoto	37	11	Epitácio Lindolfo da Silva Pessoa	30	13
Juarez do Nascimento Fernandes Távora	36	13	Roseana Macieira Sarney	30	10
Ernâni Amaral Peixoto	35	15	Ciro Ferreira Gomes	29	13

Tabela 17 - ranking com os 20 verbetes com mais menções a vínculos familiares

A quantidade de menções não deve ser tomada rigorosamente. No verbete de Getúlio Vargas, por exemplo, verificamos que nenhuma das 11 ocorrências de *companheiro* denotava qualquer tipo de relação de parentesco, mas sim de correligionário, parceiro ou coparticipante de algum evento ou empreitada. E em Raul Jungmann, das 35 relações identificadas, 28 (ou seja 80%) referem-se à *família*, mas todas fora do contexto de parentesco.

Há uma explicação para isso: os cargos que Jungmann ocupou na sua trajetória política incluem a presidência do INCRA e os ministérios da política fundiária e do desenvolvimento agrário, onde a questão dos assentamentos de *famílias* estava sempre em pauta.

Outro ponto que chama a atenção é que nem sempre um alto número de ocorrências significa também alta diversidade de relações. Um exemplo disso é o verbete de Luis Inácio Lula da Silva, que apesar de ter praticamente a mesma frequência de Eurico Gaspar Dutra, este consegue ter uma diversidade quase duas vezes maior que Lula.:

Verbete	Total Freq.	Tipo	Qtd	Verbete	Total Freq.	Tipo	Qtd
Luís Inácio da Silva	41	compadre	2	Eurico Gaspar Dutra	40	casal	1
		família	12			companheiro	1
		filha	2			enteada	1
		filho	3			enteado	1
		irmão	7			esposa	1
		mãe	9			família	1
		pai	4			filha	1
		prima	1			filho	9
		viúvo	1			genro	2

						irmão	8
						marido	1
						mãe	1
						pai	6
						parente	3
						progenitor	1
						tio	1
						viúvo	1

Tabela 18 - comparação do número de relações familiares existentes nos verbetes de Luis Inácio da Silva e Eurico Gaspar Dutra

No caso de Lula, a concentração de ocorrências nos lemas *família* e *mãe* se explica pelo destaque que o autor decidiu dar às origens do personagem, resgatando a história familiar pautada pelas dificuldades de sobrevivência e o papel fundamental da mãe no sustento dele e seus sete irmãos. Por isso há tantas menções a estes dois termos. Já no caso de Dutra, a história é diferente: observando de perto o texto do verbete, vemos que das 40 ocorrências identificadas, 18 referem-se a laços familiares de terceiros e 10 não têm a ver diretamente com vínculos de parentesco. Ou seja, apenas 12 (30% do total) estão circunscritos ao seu próprio círculo familiar.

O viés quantitativo sempre vai conter ruídos e deve ser tomado com ressalvas. Por outro lado, ainda que não sejam fornecidas respostas precisas, a identificação destas relações é um guia em potencial por onde podemos começar a olhar, quando o assunto são vínculos familiares na política. Veremos no próximo capítulo, os padrões textuais para a extração destas relações no DHBB e os resultados obtidos.

A metodologia mostrada nesta seção pode ser utilizada para anotar informação sobre qualquer outro campo semântico de interesse no corpus.

5.4
Etapa 4: Aplicação

A ideia nesta etapa é selecionar alguns temas específicos, como nascimento, formação acadêmica e vínculos familiares, e identificar um conjunto de padrões que possam ser testados e avaliados quanto à sua produtividade em relação ao DHBB.

Podemos resumir o processo da seguinte forma: para cada tema, observamos em uma amostra de verbetes como se dão as construções das frases que trazem a informação desejada e separamos o maior número possível delas, considerando diversidade e abrangência.

Em seguida traduzimos estas construções em padrões léxico-sintáticos combinando expressões regulares e marcações *POS* do analisador adotado, testando iterativamente os filtros até obtermos as frases que nos interessam.

Concatenamos todas as expressões e aplicamos diretamente no sistema AC/DC da Linguateca por meio do terminal CQP (Corpus Query Processor). O resultado, como veremos na etapa de resultados, é levado para o programa R, onde os trechos de interesse são isolados e sintetizados em um *dataframe* e as informações pontuais extraídas com regras mais finas, para então serem cruzadas com outros metadados.

5.4.1
Amostra

Apesar de os padrões recuperarem informações de todas as biografias existentes no DHBB, elegemos uma amostra que permitisse a checagem manual dos resultados extraídos para alguns dos exercícios propostos. Nessa amostra colocamos os presidentes da República, os ministros de Estado, os senadores e os deputados federais, todos ocupantes de cargos dos poderes Executivo e Legislativo na esfera federal. Dessa forma conseguimos medir com mais acuidade o grau de confiança desses resultados, além de obter um perfil representativo de cada uma dessas classes.

Os grupos eleitos estão discriminados na tabela a seguir, com a ressalva de que um indivíduo pode ter ocupado mais de um cargo ao longo da sua trajetória política, comparecendo assim em mais de um grupo:

Cargo	Gênero	Total de indivíduos	
Presidentes da República	f	0	26
	m	26	
Ministros	f	27	894
	m	867	

Senadores	f	24	742
	m	718	
Deputados federais	f	160	4.314
	m	4.154	

Tabela 19- Seleção dos cargos para amostra a título de avaliação dos exercícios

Descartando as duplicidades, o número total de indivíduos únicos é de 5.219. A título de esclarecimento, quando esta tese está sendo escrita, muitos políticos do cenário atual ainda não receberam atualizações ou sequer foram incluídos no DHBB, o que pode levar à falsa impressão de que os números acima estão incorretos. Dos presidentes da República, por exemplo, constam desde Epitácio Pessoa (que governou entre 1919 e 1922) até Luis Inácio da Silva (2003-2011). Dilma Rousseff, Michel Temer e Jair Bolsonaro, últimos a exercerem a posição, não estão listados no grupo dos Presidentes porque não tiveram seus verbetes atualizados até então. Mas comparecerem no exercício de outros cargos. A versão que utilizamos do DHBB na Linguateca para obter estes números foi a 7.3[41].

[41] https://www.linguateca.pt/acesso/desc_dhbb.html. Simulações realizadas a partir da versão utilizada podem trazer resultados diferentes.

5.4.2
Temas

Os temas abaixo foram criados a partir das perguntas enviadas pelos pesquisadores interessados em consultar o DHBB[42], e serão usados para testar a metodologia. A incorporação dos metadados (gênero e cargos ocupados) aos resultados foram particularmente úteis, permitindo o cruzamento com as informações extraídas.

1. **Com que idade o indivíduo iniciou sua carreira pública?**
   - Extração dos trechos contendo informação sobre nascimento;
   - Metadados contendo as linhas de cargos e respectivos períodos de cada exercício;
2. **Qual a formação acadêmica do político?**
   - Extração dos trechos contendo informação sobre formação acadêmica;
3. **O que dizer sobre os vínculos familiares na política?**
   - Extração dos trechos contendo informação sobre vínculos familiares com outros personagens também políticos (isto é, outros 'verbetados').

Certamente a maioria desses dados são encontrados atualmente na Internet, em sites como os da Câmara[43], do Senado[44] e outras fontes não oficiais. O exercício tem o intuito de medir a recuperação destas informações no DHBB.

5.4.3
Padrões

Esta é a etapa onde: i) observamos em uma amostra de verbetes as construções das frases que trazem a informação desejada, ii) traduzimos estas construções em padrões léxico-sintáticos combinando expressões regulares e

[42] Conforme explicado na seção 1.1, sobre as principais motivações da pesquisa. A lista completa das perguntas encontra-se no Anexo 1.
[43] https://www.camara.leg.br/
[44] https://www12.senado.leg.br/hpsenado

etiquetas de anotação, iii) concatenamos todas as expressões (quando houver mais de uma) e iv) aplicamos no corpus.

Para cada tema, apresentamos algumas frases de exemplo e as expressões criadas.

Sobre dados de nascimento

De longe, é a informação mais direta de se obter pois sua escrita obedece a um certo padrão: ela é constituída da data e local do evento, localiza-se sempre no primeiro parágrafo, é precedida do nome do biografado e seguida dos nomes dos pais, quando mencionados. Criamos a expressão abaixo para abranger estes casos:

NASCIMENTO	
Frases de exemplo	«Moroni Bing Torgan» nasceu em Porto Alegre, no dia 10 de junho de 1956.
	«Álvaro Francisco de Sousa» nasceu no dia 28 de fevereiro de 1903.
Expressão	[classe="bio.*" & dicionario="dhbb" & pos="PROP.*"]+ [: pos!="PROP.*" :][]{0,1} [lema="nascer" & word!="nascido|nascer"] [pos="PRP.*"][]{0,21} [pos="NUM.*|ADJ.*"] [word="de"]? [pos="N.*"]? [word="de"]? [pos="NUM.*"]? [pos="PU"]
Ocorrências	6.464

Basicamente, a sintaxe quer dizer: "faça apenas para verbetes biográficos do DHBB, encontre todas as construções onde exista um [nome próprio composto], sucedido do [lema 'nascer'] (que não pode ser 'nascido' nem 'nascer') e de uma [preposição]. Siga até encontrar o primeiro [número] ou [modificador] da sentença (dia, ano ou mês) e continue até a próxima vírgula ou ponto final; neste intervalo pode ou não haver outro [número] (eventualmente, a outra parte da data)".

Abrindo o terminal CQP (Corpus Query Processor), executamos o script que acessa o corpus, aplicamos a expressão de busca e gravamos todas as ocorrências que se encaixam nela em um arquivo que nomeamos como cqpNascimento.txt.

342: <texto id=dhbb10 raw>: « <Armando Abílio Vieira » nasceu em Itaporanga (PB) no dia 29 de dezembro de 1944 ,> filho de Argemiro Abílio de Sousa e de Luísa Bronzeado Vieira .

1974: <texto id=dhbb100 raw>: « <Pedro Aleixo » nasceu em São Caetano , distrito do município de Mariana (MG) , no dia 1º de agosto de 1901 ,> filho do comerciante José Caetano Aleixo e de Úrsula Martins Aleixo .

9024: <texto id=dhbb1000 raw>: « <Eduardo Henrique Accioly Campos » nasceu em Recife , no dia 10 de agosto de 1965 ,> filho de Maximiano Accioly Campos e de Ana Lúcia Arraes de Alencar .

Figura 8 - Excerto do arquivo com ocorrências para o padrão de nascimento

Cada ocorrência mostra a linha em que ela aparece no corpus, o ID do verbete e entre os sinais "<" e ">" o trecho que supostamente estamos interessados.

Em paralelo, realizamos algumas operações no R: i) a partir dos arquivos originais dos verbetes, extraímos os metadados e criamos um dataframe com as seguintes colunas: ID, nome do verbete, sexo e cargos (com respectivos períodos de ocupação); ii) abrimos o arquivo gerado no CQP ("cqpNascimento.txt") e extraímos o ano de nascimento do biografado; iii) incluímos a informação no dataframe criado anteriomente; iv) identificamos a data associada ao primeiro cargo público ocupado pelo indivíduo e calculamos a idade correspondente à época.

O resultado é este dataframe:

	dhbb_id	verbete	sexo	cargos	sent	nasc	cargo_1	ano_cargo_1	idade_cargo_1
1	4463	Aarão Rabelo	m	@ const. 1934 @ dep. fed. SC 1962	Aarão Rabelo nasceu em Itajaí (SC) no dia 26 de fevereir...	1906	@ const. 1934	1934	28
2	5204	Aarão Steinbruch	m	@ dep. fed. RJ 1953 @ dep. fed. RJ 1955-1963 @ sen. ...	Aarão Steinbruch nasceu em Santa Maria (RS) em 17 de...	1917	@ dep. fed. RJ 1953	1953	36
3	3750	Abdias do Nascimento	m	@ dep. fed. RJ 1983-1986 @ sen. RJ 1991-1992 @ sen. ...	Abdias do Nascimento nasceu em Franca (SP) no dia 14...	1914	@ dep. fed. RJ 1983	1983	69
4	2379	Abdon Gonçalves Nanhay	m	@ dep. fed. RJ 1975-1979	Abdon Gonçalves Nanhay nasceu em São João de Meriti...	1928	@ dep. fed. RJ 1975	1975	47
5	4945	Abdon Sena	m	@ militar @ comte. Comdo. Mil. Brasília 1966-1968	Abdon Sena nasceu em Florianópolis no dia 10 de julho...	1912	@ comte. Comdo. Mil. Brasília 1966	1966	54
6	411	Abel Ávila dos Santos	m	@ dep. fed. SC 1971-1979	Abel Ávila dos Santos nasceu em Tijucas (SC) no dia 18 ...	1917	@ dep. fed. SC 1971	1971	54
7	1340	Abel de Abreu Chermont	m	@ dep. fed. PA 1918-1920 @ rev. 1930 @ junta gov. PA ...	Abel de Abreu Chermont nasceu em Belém no dia 21 de ...	1887	@ dep. fed. PA 1918	1918	31
8	2910	Abel dos Santos Lima	m	@ dep. fed. CE 1961-1962	Abel dos Santos Lima nasceu em Lavras da Mangabeira (...	1912	@ dep. fed. CE 1961	1961	49
9	4852	Abel José dos Santos	m	@ dep. fed. prof. 1935-1937	NA	NA	@ dep. fed. prof. 1935	1935	NA
10	3117	Abel Pinheiro Maciel Filho	m	@ gov. AC 1953-1954	Abel Pinheiro Maciel Filho nasceu no vale do Juruá , est...	1895	@ gov. AC 1953	1953	58
11	4420	Abel Rafael Pinto	m	@ dep. fed. MG 1959-1967 @ dep. fed. MG 1969	Abel Rafael Pinto nasceu em Paraíba do Sul (RJ) no dia ...	1914	@ dep. fed. MG 1959	1959	45
12	3144	Abel Sauerbronn de Azevedo Magalhães	m	@ magistrado @ interv. RJ 1945-1946	Abel Sauerbronn de Azevedo Magalhães nasceu em Can...	1881	@ interv. RJ 1945	1945	64
13	4349	Abelardo Bretanha Bueno do Prado	m	@ diplomata @ encar. neg. Bras. EUA 1936-1937 @ em...	Abelardo Bretanha Bueno do Prado nasceu em Jaguarã...	1896	@ encar. neg. Bras. EUA 1936	1936	40
14	932	Abelardo Calafange	m	@ dep. fed. RN 1951-1954	Abelardo Calafange nasceu em Canguaretama (RN) no ...	1904	@ dep. fed. RN 1951	1951	47
15	2630	Abelardo de Araújo Jurema	m	@ sen. PB 1953-1954 @ sen. PB 1957 @ dep. fed. PB 19...	Abelardo de Araújo Jurema nasceu em Itabaiana (PB) n...	1914	@ sen. PB 1953	1953	39
16	3331	Abelardo dos Santos Mata	m	@ militar @ const. 1946 @ dep. fed. RJ 1946-1955	Abelardo dos Santos Mata nasceu no Rio de Janeiro , e...	1906	@ const. 1946	1946	40
17	279	Abelardo Fortuna Andréia dos Santos	m	@ militar @ dep. fed. BA 1951-1955	Abelardo Fortuna Andréia dos Santos nasceu em Salvad...	1912	@ dep. fed. BA 1951	1951	39
18	1426	Abelardo Leão Conduru	m	@ sen. PA 1935-1937	Abelardo Leão Conduru nasceu em Belém no dia 17 de f...	1889	@ sen. PA 1935	1935	46
19	3039	Abelardo Luís Lunion Melo	m	@ dep. fed. PR 1992 @ dep. fed. PR 1994-1995 @ dep.	Abelardo Luís Lunion Melo nasceu em Curitiba no dia ?	1953	@ dep. fed. PR 1992	1992	40

Figura 9 - Sistematização dos dados extraídos

Com as informações consolidadas, conseguimos agora fazer cruzamentos com os metadados e separar as ocorrências por categorias (gênero, cargos, idade etc.), conforme veremos no próximo capítulo.

Sobre formação acadêmica

Informações sobre a formação acadêmica, quando existentes, localizam-se nos verbetes biográficos logo após o parágrafo inicial sobre nascimento e filiação do titular, dispostas em ordem cronológica. A estrutura de enunciação é mais complexa que a de nascimento e pode envolver uma diversidade grande de verbos e complementos, além de um número sem fim de construções: "Graduado em economia pela Faculdade de Ciências Econômicas da Universidade Federal do Ceará (UFC), concluiu o curso de engenharia eletrônica no Instituto de Tecnologia da Aeronáutica"; "Ingressou na Faculdade de Direito da Universidade de Minas Gerais (UMG)".

Como no exercício anterior, as regras foram criadas após a observação de uma amostra de verbetes. A partir dela, reunimos o maior número de padrões que abrangesse as construções possíveis e convertemos em expressões combinadas. No total foram 11 expressões criadas[45], uma delas exemplificada abaixo:

FORMAÇÃO	
Frases de exemplo	se graduou em ciências sociais na Universidade de São Paulo (USP).
	diplomou-se como engenheiro naval em 1964, na Escola Nacional de Engenharia.
Sintaxe de busca	[word="se"]? [dicionario="dhbb" & classe="biográfico" & lema="especializar.*|diplomar.*|bacharelar.*|graduar.*|doutorar.*" & word!= "especializad.?s|especializada"] []* "\."
Ocorrências	2.890

A sintaxe pode ser interpretada da seguinte forma: "encontre todas as construções onde [pode ou não existir a partícula 'se'] [seguido de um dos lemas: 'especializar', 'bacharelar', 'graduar' ou 'doutorar'] e traga tudo o que vier em seguida até o ponto final. Note ainda que há uma regra para restringir a ocorrência das palavras 'especializada', 'especializadas' e 'especializados', a fim de evitar construções como: "Diplomado em guarda-livros, trabalhou na firma M. I. Castro, especializada em caixas de madeira para exportação de borracha e castanhas".

[45] As expressões na íntegra encontram-se no Anexo 3 da tese.

Da mesma forma que no exercício anterior, aplicamos as expressões no corpus a partir do terminal CQP, recuperando 10.565 trechos relativos à formação dos biografados no DHBB.

A maioria dos verbetes (2.917) recupera apenas um trecho, mas muitos também trazem dois ou mais:

Trechos recuperados	1x	2x	3x	4x	5x	6x	7x	8x	9x	10x	11x	13x
por verbete	2.917	1.476	683	314	126	64	23	11	6	5	1	1
Total	10.565 trechos											

Tabela 20 - Extração de informações sobre formação acadêmica nos verbetes biográficos

Observando de perto o cruzamento entre a lista de frequência, os trechos recuperados e os cargos ocupados pelos titulares, percebemos que as biografias que trazem maior número de ocorrências são aquelas que, salvo algumas poucas exceções (em que o autor incluiu informações de ensino primário e secundário), a grande maioria diz respeito aos militares, ou seja, há uma tendência em descrever com detalhes a formação destes, enumerando todos os cursos preparatórios, de aperfeiçoamento e de comando que realizaram.

Os 10.565 trechos foram recuperados de 5.627 verbetes, que representam cerca de 83% do total de verbetes biográficos do DHBB.

Novamente, realizamos algumas operações no R: i) a partir dos arquivos originais dos verbetes, extraímos os metadados e criamos um dataframe com as

dhbb_id	verbete	sexo	cargos	trechos
1010	Jaime Mendonça de Campos	m	@ dep. fed. RJ 1989-1991	Formado em direito pela Universidade Federal do Rio de Ja...
1010	Jaime Mendonça de Campos	m	@ dep. fed. RJ 1989-1991	realizou diversos cursos na área de direito , quase todos na ...
1011	João Elísio Ferraz de Campos	m	@ gov. PR 1986-1987	Bacharel em direito em 1966 pela Pontifícia Universidade Ca...
1012	José Eduardo Siqueira Campos	m	@ dep. fed. TO 1989-1992 @ sen. TO 1999-	estudou no Centro de Ensino Unificado de Brasília (CEUB) , ...
1013	Júlio José Campos	m	@ dep. fed. MT 1979-1983 @ gov. MT 1983-1986 @ const...	estudou em Goiás onde fundou e presidiu a Associação do E...
1013	Júlio José Campos	m	@ dep. fed. MT 1979-1983 @ gov. MT 1983-1986 @ const...	Cursou agronomia na Universidade Estadual Paulista Júlio d...
1014	Lauro Campos	m	@ sen. DF 1995-2003	Formado em direito pela Universidade Federal de Minas Ger...

Figura 10 - Dataframe com os trechos extraídos

colunas ID, nome do verbete, sexo e cargos, abrimos o arquivo gerado no CQP (“cqpFormacao.txt”) e extraímos as áreas de formação com apoio adicional de um léxico de áreas.

Muitas formações têm uma mesma raiz, variando apenas nas suas especializações ou na forma como é referenciada nos textos. A fim de melhorar o resultado das frequências criamos uma lista de correspondências e aplicamos aos trechos recuperados. Dessa forma, "ciências jurídicas" e "advogado", por exemplo, são incluídos em "direito"; "sociologia" em "ciências-sociais"; "cardiologista" em "medicina", e assim por diante. No final obtivemos uma visão mais enxuta das áreas percorridas pelos políticos.

O resultado é esta tabela contendo o número de ocorrências das áreas de formação presentes nas biografias:

Áreas	Ocorrências	Áreas	Ocorrências
direito	3851	relações-internacionais	70
formação-militar	1082	políticas-públicas	67
engenharia	912	geografia	66
medicina	839	psicologia	56
economia	633	criminologia	39
administração	613	educação-física	39
filosofia	289	arquitetura	36
contabilidade	192	veterinária	32
letras	190	estatística	22
ciências-sociais	147	ciências-naturais	19
teologia	133	geologia	19
agronomia	121	informática	18
educação	114	serviço-social	18
história	114	metalurgia	14
humanidades	100	urbanismo	14
ciências-políticas	86	eletrotécnica	13
química	78	turismo	12
pedagogia	77	biologia	11
jornalismo	76	radiologia	11

farmácia	73	belas-artes	8
física	72	cinema	7
comunicação	71	enfermagem	4
odontologia	71	nutrição	4
matemática	70	biblioteconomia	1

Tabela 21- Áreas de formação dos biografados

No total, foram compiladas 48 formações acadêmicas presentes nas sentenças recuperadas. No capítulo 6 veremos como se apresenta a distribuição das áreas entre os políticos, de acordo com as gerações e cargos ocupados.

Sobre vínculos familiares

Na seção 5.3.6 mostramos como foi feito o enriquecimento semântico do corpus a partir de anotações sobre vínculos familiares. O enriquecimento simplificou e tornou mais produtiva a extração de informações deste campo semântico.

Como nos exercícios anteriores, as regras foram criadas após a observação de uma amostra de verbetes, porém desta vez a amostra não foi aleatória. A fim de melhorar a qualidade das ocorrências, buscamos por padrões em uma seleção de verbetes cujos titulares já são conhecidos por terem vínculos familiares com outros políticos[46].

Desta seleção, os dez verbetes com maior número de menções de parentesco no DHBB[47] são: José Ribamar Ferreira de Araújo Costa (Sarney), Fernando Afonso Collor de Melo, Antônio Carlos Ribeiro de Andrada, Ciro Ferreira Gomes, José Renan Vasconcelos Calheiros, Jáder Fontenelle Barbalho, Antônio Carlos Peixoto de Magalhães, Tancredo de Almeida Neves, Miguel Arrais de Alencar, Ernâni Amaral Peixoto.

[46] A edição da Folha de S. Paulo de 19 de agosto de 2018 traz uma matéria sobre as dinastias políticas no Brasil e lista as principais famílias que comandam ou comandaram a política em determinados estados e regiões do país. Ver no Anexo 5.
[47] Sobre o campo semântico relações familiares, consulte a seção 5.3.6.

O primeiro passo foi procurar por sentenças onde laços familiares ocorrem. Os termos ‘companheiro’ e ‘família’ foram ignorados, pois apesar de pertencerem ao léxico familiar, no DHBB aparecem apenas em relações como “*companheiro* de chapa”, “administrador das empresas da *família*” ou “auxílio bolsa *família*”. Usamos a seguinte expressão:

[dicionario="dhbb" & fonte="José Ribamar Ferreira de Araújo Costa|Fernando Afonso Collor de Melo|Antônio Carlos Ribeiro de Andrada|Ciro Ferreira Gomes|José Renan Vasconcelos Calheiros|Jáder Fontenelle Barbalho|Antônio Carlos Peixoto de Magalhães|Tancredo de Almeida Neves|Miguel Arrais de Alencar|Ernâni Amaral Peixoto" & sema="familia:lacos.*" & word!="companheiro.*|família.*"]

No total, foram recuperadas 221 ocorrências. Olhando de perto para cada uma delas, foi possível classificá-las em relações do tipo válidas e relações não válidas, partindo tanto de uma suposta relação familiar do biografado com outras pessoas, quanto de relações de terceiras pessoas mencionadas nos verbetes.

Em geral, as relações válidas para o biografado acontecem por meio de pronomes anafóricos, que, como vimos no capítulo 4.3.2, representam cerca de 46% das orações do DHBB. As relações que consideramos não válidas dizem respeito aos casos em que o termo de família não representa um vínculo direto com ninguém (ou seja, encontra-se ‘solto’ na sentença) ou não demonstra nenhuma relação familiar.

Tipo		Exemplos	N.	%
Relações válidas para o biografado	Com outra pessoa com nome próprio	*No Ceará,* ***seu irmão Cid Gomes****, também do PSB, foi eleito governador do estado ainda no primeiro turno*	95	(132) 59,7%
	Com outra pessoa, sem nome próprio	***Seu pai foi governador*** *(1951-1956) e senador por Alagoas (1963-1981)*	13	
Relações válidas entre terceiros	Entre pessoas com nome próprio	*deixou profundas mágoas em* ***Carlos Francisco, irmão de Tasso*** *e administrador das empresas da família*	24	
Relações não válidas	Relação com outra pessoa mencionada distante	*No mesmo dia 12, Vargas estava em Belo Horizonte junto com Tancredo Neves para participar da inauguração dos novos fornos da Siderúrgica Mannesmann. Depois de proferir um discurso enérgico em defesa da legalidade, que considerava*	12	(89) 40,3%

		ameaçada, retornou ao Rio, onde tomou conhecimento de que ***seu filho, o deputado federal Lutero Vargas****...* [Verbete: Tancredo Neves]		
	Relação indireta	*transferiu-se para Maceió e assumiu a direção da «Gazeta de Alagoas», jornal de propriedade de* ***seu pai.***	71	
	Não são relações	*alimentação de gestantes, de jovens* ***mães*** *e de crianças;*	6	

Tabela 22 - Relações válidas e não válidas das relações familiares obtidas da amostra

Das 221 ocorrências, apenas 132 são relações válidas, e dentre estas, 108 representam vínculos do biografado com outro indivíduo (identificado principalmente, como já dito, por um pronome anafórico), podendo ou não ser referenciado por um nome próprio ou pelo cargo/papel que a pessoa ocupa. As outras 24 são relações entre terceiras pessoas mencionadas no verbete. Das 89 ocorrências que não estão sendo consideradas válidas, a mais difícil de identificar talvez seja a relação que liga a uma terceira pessoa mencionada anteriormente no texto, distante do termo.

Em seguida, partimos das sentenças válidas para encontrar semelhanças e hipotetizar padrões que indiquem as relações familiares que nos interessam.

Padrão 1: <nome próprio>, <família:laço>

- [Jorge Murad], [ex-genro]

Padrão 2: <família:laço> de <nome próprio>

- [neto] de [Martim Francisco Ribeiro de Andrada]

Padrão 3: <família:laço> do <cargo/papel> <nome próprio>

- [filho] do [senador] [Benedito de Lira]

Padrão 4: seu <família:laço> foi <cargo/papel>

- Seu [pai] foi [governador]

Padrão 5: seu <família:laço>, <nome próprio>

- Seu [avô materno], [Lindolfo Collor]

Padrão 6: seu <família:laço> <nome próprio> foi <cargo/papel>

- Seu irmão Maurício Graco Cardoso foi deputado

Padrão 7: o <cargo/papel> <nome próprio>, <família:laço> de <nome próprio>

- o [deputado federal] [Renan Filho], [filho] de [Renan Calheiros],

Padrão 8: <família:laço>. Entre eles, <nome próprio>, <cargo/papel>

- [filhos]. Entre eles, [Roseana Sarney], [deputada federal] pelo Maranhão

Padrão 9: <nome próprio> e seu <família:laço> <nome próprio>

- Amaral Peixoto e seu genro, Wellington Moreira Franco

Padrão 10: <nome próprio>, <família:laço> de <nome próprio>

- [Maria Fernanda Quintela Brandão Vilela], [filha] de [Teotônio Vilela]

Padrão 11: sua <família:laço> era <família:laço> de <nome próprio>

Há muitas variações de escrita encontradas no corpus, mesmo seguindo estes padrões. Derivamos no total 33 expressões [48]: 13 delas recuperam relações

[48] As expressões na íntegra encontram-se no Anexo 3 da tese.

familiares do biografado com outras pessoas e 10 recuperam relações entre terceiros. Abaixo, um exemplo:

LAÇOS FAMILIARES NO DHBB	
Frase de exemplo	Seu filho, Fernando Ramos de Alencar, sobressaiu-se na carreira diplomática
Expressão	[dicionario="dhbb" & pos="DET.*"][sema="familia:lacos.*" & word!="companheiro.*|família.*"] ","? [entidade="dhbb.*"]+ [: entidade!="dhbb.*" :]
Ocorrências	655

A sintaxe significa: “encontre todas as construções do DHBB que [inicia com um pronome] [seguido de um laço familiar] [podendo ou não ser seguido de uma vírgula] [seguido do nome de um biografado]”.

Aplicamos as expressões em todo o corpus, separando pelas categorias válidas de relação. Aproveitando o fato de que é possível discriminar os biografados do DHBB nas expressões de busca, criamos uma categoria especialmente para identificá-los. Os números encontrados estão discriminados na tabela abaixo

Tipo		N. Expressões	N. ocorrências	%
Relações válidas para o biografado	(1) Com outro biografado do DHBB	8	1.194	94 %
	(2) Com outra pessoa com nome próprio	11	8.796	
	(3) Com outra pessoa sem nome próprio	4	624	
Relações válidas entre terceiros	(4) Entre biografados do DHBB	3	35	6 %
	(5) Entre pessoas com nome próprio	7	640	

Tabela 23 - Relações válidas das relações familiares obtidas em todo o DHBB

Das relações acima, é preciso destacar que as relações (1) certamente estão contidas em (2) e as de (4) estão em (5), tendo sido separadas apenas com o intuito de demarcar os vínculos mencionados entre os biografados que constam no DHBB.

A maioria dos vínculos familiares recuperados dizem respeito ao próprio biografado com outro indivíduo (podendo ser um biografado, uma pessoa identificada pelo nome próprio ou pelo cargo/papel). Menos de 10% representam relações entre terceiras pessoas mencionadas no verbete.

5.5
Etapa 5: Resultados

Nesta seção descrevemos os resultados obtidos nas extrações realizadas para cada um dos temas propostos e a avaliação conduzida em cima destes resultados.

5.5.1
Métricas

Precisão e abrangência são medidas amplamente utilizadas para avaliar a qualidade dos resultados em áreas do conhecimento que lidam com reconhecimento de padrões e recuperação de informações. Precisão é uma medida de relevância, enquanto abrangência é uma medida de cobertura e ambas são obtidas a partir de cálculos baseados na avaliação do que foi recuperado e o que deixou de ser recuperado (verdadeiros-positivos, verdadeiros-negativos, falsos-positivos, falsos-negativos). Uma matriz de confusão com esta classificação nos permitirá encontrar a medida-F (F1), que é a média harmônica ponderada de precisão e abrangência do nosso modelo.

	Relevante	Não relevante
Recuperado	Verdadeiro-positivo (VP)	Falso-positivo (FP)
Não recuperado	Falso-negativo (FN)	Verdadeiro-negativo (VN)

Tabela 24 - Matriz de confusão entre informações extraídas x informações desejadas

Métricas de avaliação:

$$\text{Precisão: } \frac{\text{Itens relevantes recuperados}}{\text{Total de itens recuperados}} \qquad P = \frac{VP}{(VP + FP)}$$

$$\text{Abrangência: } \frac{\text{Itens relevantes recuperados}}{\text{Total de itens relevantes}} \qquad R = \frac{VP}{(VP + FN)}$$

$$\text{Medida-F: } \frac{2 * (\text{Precisão} * \text{Abrangência})}{(\text{Precisão} + \text{Abrangência})} \qquad F1 = \frac{2 * (P * R)}{(P + R)}$$

Logicamente, é possível aumentar a abrangência retornando mais ocorrências (isto é, tornando mais gerais as expressões de busca) ou obter altos níveis de precisão com baixos níveis de abrangência. O controle sobre as expressões, com os devidos cuidados, parece ser uma das vantagens da abordagem

via padrões. Em síntese, quanto maior a Medida-F, melhor o ponto de equilíbrio entre as medidas de precisão e abrangência.

5.5.2
Avaliação

Sobre dados de nascimento

A extração dos dados de nascimento nos trouxe os seguintes dados que consolidamos na tabela abaixo.

Ocorrências	Abrangência e precisão na recuperação de sentenças com informações sobre nascimento														
	Todos verbetes biográficos do DHBB			Apenas presidentes República			Apenas ministros			Apenas senadores			Apenas deputados federais		
Total de verbetes	**6.745 (100%)**			**26 (100%)**			**894 (100%)**			**742 (100%)**			**4.314 (100%)**		
Verbetes recuperados	6.455 **95,7%**	VP	6.444	25 **96,1%**	VP	25	877 **98%**	VP	875	726 **97,8%**	VP	725	4.176 **96,8%**	VP	4.165
		FP	11		FP	0		FP	2		FP	1		FP	11
Verbetes não recuperados	290 **4,3%**	VN	235	1 **3,9%**	VN	0	17 **2%**	VN	4	16 **2,2%**	VN	3	138 **(3,2%)**	VN	68
		FN	54		FN	1		FN	13		FN	13		FN	70
Precisão	99%			100%			99%			99%			99%		
Abrangência	99%			96%			98%			98%			98%		
Medida-F	99%			97%			98%			98%			98%		

Tabela 25 - Avaliação da extração de informações sobre nascimento

Por ser uma informação bastante pontual - apenas ano de nascimento -, foi possível checar manualmente todas as ocorrências. No geral, a taxa de acertos na recuperação da data ficou em torno de 98%, já considerando os verbetes não recuperados pela própria ausência dessa informação no texto (verdadeiros-negativos).

O método de checagem foi a seguinte: a) primeiro, verificamos quais verbetes capturaram alguma informação (“verbetes recuperados”); caso tenham trazido um ano associado que não seja divergente, são considerados corretos; caso não, verificamos na origem se de fato não consta este dado apesar do *match* na construção da sentença (então é considerado correto também) ou se deveria ter sido recuperado, mas não aconteceu por algum erro; b) verificamos quais verbetes não capturaram nada: da mesma forma, caso realmente não haja nada para recuperar

(checado manualmente) está correto, caso contrário, se tratará de erro; c) verificamos se há duplicidade de verbetes com sentenças diferentes, o que indicaria divergência de informação visto que deve existir apenas uma data de nascimento associada ao titular.

No caso dos deputados federais, que são em maior número, contabilizamos 4.176 verbetes onde trechos correspondentes ao padrão léxico-sintático utilizado foram detectados. Como o total de deputados existentes no DHBB é de 4.314, esse número corresponde a 96,8% do total. Destes, 11 estão incorretos, seja porque a parte contendo as informações da data de nascimento não foi recuperada (ainda que ela esteja presente no texto) ou porque a ocorrência dizia respeito a outro trecho do verbete que também "casava" com a sintaxe (aconteceu em um único caso). Já 138 deputados, ou seja 3,2% do total, não compareceram nos resultados e neste caso temos quase um empate entre erros e acertos: 49% destas ausências se justificam pois não há nenhum dado de nascimento fornecido sobre o titular do verbete, então está correto, e 51% delas decorre ou porque a construção da sentença foge ao padrão léxico-sintático utilizado ou porque o parser falhou na anotação.

Sobre formação acadêmica

Os 10.565 trechos sobre formação acadêmica foram recuperados de 5.627 verbetes, que representam cerca de 83% do total de verbetes biográficos do DHBB.

Cada uma das ocorrências foi manualmente checada para identificação dos casos em que a construção léxico-sintática tenha atendido a um dos padrões definidos, porém a informação contida ali não é válida. Se a sentença extraída dizia respeito a cursos realizados, títulos obtidos, ingresso em universidades e eventos afins, então ela está válida, caso contrário, não. Nem todas trazem explicitamente a área de formação ou nível de escolaridade que o indivíduo adquiriu, mas toda informação neste sentido fornece pistas para a construção de seu perfil educacional.

Total de ocorrências	10.565 (100%)
Ocorrências válidas	10.470 (99,1%)
Ocorrências não válidas	95(0,9%)

Tabela 26 – Avaliação da extração de informações sobre formação

Exemplos de casos não válidos:

"especializado em notícias policiais, sendo demitido dois meses depois ..."

"formado no processo constituinte, reunindo parlamentares de centro-esquerda ..."

"graduados da Marinha e da Aeronáutica que se haviam rebelado na capital federal..."

Não é possível saber quantas e quais ocorrências deixaram de ser recuperadas em todo o corpus, por isso a avaliação não contempla medidas de precisão ou abrangência. Mas o índice de acertos de 99,1% para as ocorrências válidas das informações extraídas mostra-se promissor.

Sobre vínculos familiares

Para fins de avaliação, as expressões criadas para extração de vínculos familiares no DHBB foram aplicadas na mesma amostra de verbetes que consideramos produtivos em termos de relações de parentesco na política: José Ribamar Ferreira de Araújo Costa (Sarney), Fernando Afonso Collor de Melo, Antônio Carlos Ribeiro de Andrada, Ciro Ferreira Gomes, José Renan Vasconcelos Calheiros, Jáder Fontenelle Barbalho, Antônio Carlos Peixoto de Magalhães, Tancredo de Almeida Neves, Miguel Arrais de Alencar e Ernâni Amaral Peixoto.

Para lembrar, foram identificadas anteriormente neste conjunto 221 menções a termos de família, sendo que apenas 132 foram consideradas parte de relações válidas (seção 5.4.3). Como todas as menções foram checadas manualmente, servirão como "gabarito" para avaliar o resultado da extração, ou seja, observar o que ficou de fora, o que não deveria ter sido recuperado e as possíveis falhas na elaboração dos padrões.

A seguir, a consolidação do resultado da extração na amostra.

Ocorrências	Amostra de 10 verbetes (mais produtivos)		
Total de relações presentes na amostra (válidas e não válidas)	221 (+ 4 duplicatas)		
Relações recuperadas	**113** **50,2%**	VP	94
		FP	19
Relações não recuperadas	112 **49,8%**	VN	97
		FN	15

Precisão	83%
Abrangência	86%
Medida-F	84%

Tabela 27 - Avaliação da extração de relações familiares

Abordagens de extração por regras costumam ter taxas de precisão consideradas altas, pois as regras são construídas manualmente analisando-se atentamente os contextos léxico-sintáticos onde a informação desejada se encontra. O resultado deste exercício não foi diferente, e claro que o fato de os padrões terem sido criados a partir dessa mesma amostra tem influência no resultado, porém é interessante observar os casos que não foram bem-sucedidos.

Como já havíamos revisado todas as ocorrências anteriormente, fica fácil fazer a verificação pois se trata do mesmo conjunto de trechos. As 33 expressões utilizadas recuperaram 113 ocorrências, que representa praticamente metade de todas as relações existentes (válidas e não válidas). Destas, 94 estão corretas (83%) e as outras 19 não são válidas (são casos relacionados a problemas de correferência ou ambiguidade). Quanto às relações que não foram recuperadas, os padrões também funcionaram bem, deixando de capturar na sua maior parte aquilo que não deveria mesmo. Apenas 15 escaparam das expressões utilizadas.

Já havíamos previsto que o problema de correferenciação seria difícil de resolver com esta abordagem. Dos 19 casos de falsos positivos, 11 são decorrentes da presença de expressões anafóricas, como no exemplo a seguir:

> dhbb2352: Uma ação de busca e apreensão realizada pela Polícia Federal numa empresa de sua propriedade e de **seu marido, Jorge Murad**, revelou uma suposta participação do casal em irregularidades na extinta Superintendência do Desenvolvimento da Amazônia (Sudam).

O verbete em questão é do «Ciro Ferreira Gomes» e a relação “seu marido, Jorge Murad” diz respeito a uma terceira pessoa (Roseana Sarney), que está sendo mencionada no parágrafo anterior.

Houve quatro casos de duplicatas, onde a mesma relação foi extraída de duas formas diferentes, uma delas incorretamente:

> dhbb3807 : Nesse ponto, as discussões se generalizaram com a participação de **Alzira Vargas do Amaral Peixoto, filha do presidente**, Danton Coelho, ex-ministro do Trabalho, e outras pessoas presentes no salão em que o ministério se reunia .
>
> dhbb3807 : Nesse ponto, as discussões se generalizaram com a participação de Alzira Vargas do Amaral Peixoto, **filha do presidente, Danton Coelho**, ex-ministro do Trabalho, e outras pessoas presentes no salão em que o ministério se reunia.

E há casos que denotam uma relação válida, porém a construção textual é muito específica:

> dhbb3807: Descendia, por parte de **pai**, do comendador português José Antônio das Neves, que se estabelecera na cidade antes da independência do Brasil.

Neste exemplo acima, há uma relação entre o biografado e a pessoa de José Antônio das Neves, mas ela é intermediada por outra relação presente, com o pai do primeiro.

A grande dificuldade – para qualquer informação desejada – é encontrar o número adequado de expressões capaz de abranger o maior número possível de padrões, sem perder muito a precisão. Das 33 expressões utilizadas, 6 delas capturam menos de cinco ocorrências, o que nos faz perguntar se vale a pena mantê-las. Em compensação, uma delas sozinha identifica mais de 5 mil ocorrências no corpus todo. O fato é que pequenas variações que as construções textuais carregam para expressar determinada informação fazem com que elas se tornem cada vez mais específicas, e por outro lado a tentativa de generalizar ocasiona muitas ocorrências inúteis no resultado.

6
Interrogando o DHBB

Este capítulo é dedicado a pequenas 'leituras distantes' dos dados obtidos nas extrações.

6.1
Com que idade o político iniciou sua carreira pública?

Esta pergunta é interessante porque nos mostrará se há diferenças significativas entre as gerações no que se refere a faixa etária de entrada dos políticos na carreira pública.

A tabela 20 traz uma consolidação dos dados que o exercício de extração resultou. Só para lembrar, a informação sobre o ano de nascimento foi obtida a partir dos trechos recuperados via padrões e as informações sobre os cargos vieram dos metadados dos verbetes. Dessa forma foi possível fazer o cruzamento desses dados.

Para melhor visualizar a distribuição das idades pelos políticos, separamos os indivíduos por gerações de acordo com o seu ano de nascimento. Essa estratégia é inspirada também no trabalho de Conniff (1991).

	Média de idade ao ocupar o primeiro cargo público				
Nascidos em:	Todos os políticos	Apenas presidentes	Apenas ministros	Apenas senadores	Apenas deputados
Antes 1900 (geração 1)	(1.251)	(13)	(218)	(162)	(538)
	50 anos	**44 anos**	**50 anos**	**47 anos**	**46 anos**
1901-1920 (geração 2)	(1508)	(8)	(246)	(200)	(952)
	47 anos	**43 anos**	**49 anos**	**44 anos**	**44 anos**
1921-1940 (geração 3)	(1.552)	(2)	(196)	(188)	(1.112)
	47 anos	**39 anos**	**49 anos**	**46 anos**	**45 anos**
1941-1960 (geração 4)	(1.546)	(2)	(179)	(151)	(1.232)
	44 anos	**38 anos**	**46 anos**	**42 anos**	**43 anos**
1961-1980 (geração 5)	(354)	-	(25)	(22)	(301)
	38 anos		**42 anos**	**39 anos**	**37 anos**
1981-2000 (geração 6)	(21)	-	(1)	-	(25)
	27 anos		**34 anos**		**27 anos**
Todo período	(6.232)	(25)	(865)	(723)	(4.160)

Tabela 28- Evolução na média de idade do início de carreira dos políticos

Ainda que a representatividade das primeiras gerações seja muito superior à das gerações mais recentes em termos quantitativos, é visível que a cada década, os aspirantes à carreira política ingressam cada vez mais jovens no serviço público. Note que estamos falando de *qualquer* cargo público que justifique a sua inclusão no DHBB[49], ou seja, se a média identificada para a geração 1 dos presidentes da República é de 44 anos, não necessariamente eles chegaram por volta desta idade ao posto máximo, podendo ser outro cargo público. Por exemplo:

Biografado	Cargos ocupados	Nasc.	Primeiro cargo	Ano	Idade
Getúlio Dornelles Vargas	dep. fed. RS 1923-1926 min. Faz. 1926-1927 pres. RS 1928-1930 rev. 1930 pres. Rep. 1930-1945 const. 1946 sen. RS 1946-1949 pres. Rep. 1951-1954	1882	dep. fed. RS 1923	1923	41

Tabela 29 - Idade com a qual Getúlio Vargas iniciou sua carreira na esfera federal

Antes de ser presidente da República Getúlio Vargas ocupou vários cargos[50], tendo entrado para o Congresso Nacional aos 41 anos de idade, como deputado federal.

A diferença na média entre os que chamamos de geração 1 (nascidos antes de 1900) e os da geração 6 (nascidos após a década de 80) cai quase pela metade, isto é, se antes era mais comum iniciar a carreira pública por volta dos 50 anos, com o tempo isto foi mudando até atingir a média de 27 anos na última geração. Podemos dizer que esta constatação é a continuação de uma tendência já observada por Conniff (1991) em seu artigo que versa sobre seus estudos acerca da elite política brasileira do início do século XX:

> *À medida que o século avança, entretanto, os aspirantes à elite política assumiam cargos mais jovens. A idade média das primeiras gerações quando alcançaram seu primeiro*

[49] Para justificar a inclusão de determinado personagem no DHBB, os coordenadores definiram uma lista de cargos políticos que a pessoa tenha ocupado. A lista encontra-se no Anexo 4.
[50] Para conhecer a trajetória completa da vida de Getúlio Vargas, acesse o seu verbete em http://www.fgv.br/cpdoc/acervo/dicionarios/verbete-biografico/getulio-dornelles-vargas

> *cargo era de 55 anos; as segundas gerações 37 anos; e as terceiras, 32 anos.* (Conniff, 1991, p. 25)

O autor trabalhou com uma amostra de 250 verbetes de políticos que representavam os *major decision makers*, isto é, a elite participante do Poder Executivo nacional. Não está explícito quem exatamente seriam estes selecionados e que cargos ocuparam, mas afirma que são 93 nascidos antes de 1900, 136 entre 1901 e 1920 e 21 após 1921. Por serem homens de meia idade ou mais velhos, a política deve ter sido uma segunda carreira para as primeiras gerações.

A representação gráfica da distribuição das 6.232 biografias recuperadas na etapa anterior mostra a tendência de queda da média de idade desses indivíduos, que passam a se posicionar cada vez mais abaixo dos 40 anos, principalmente os nascidos a partir da década de 1960.

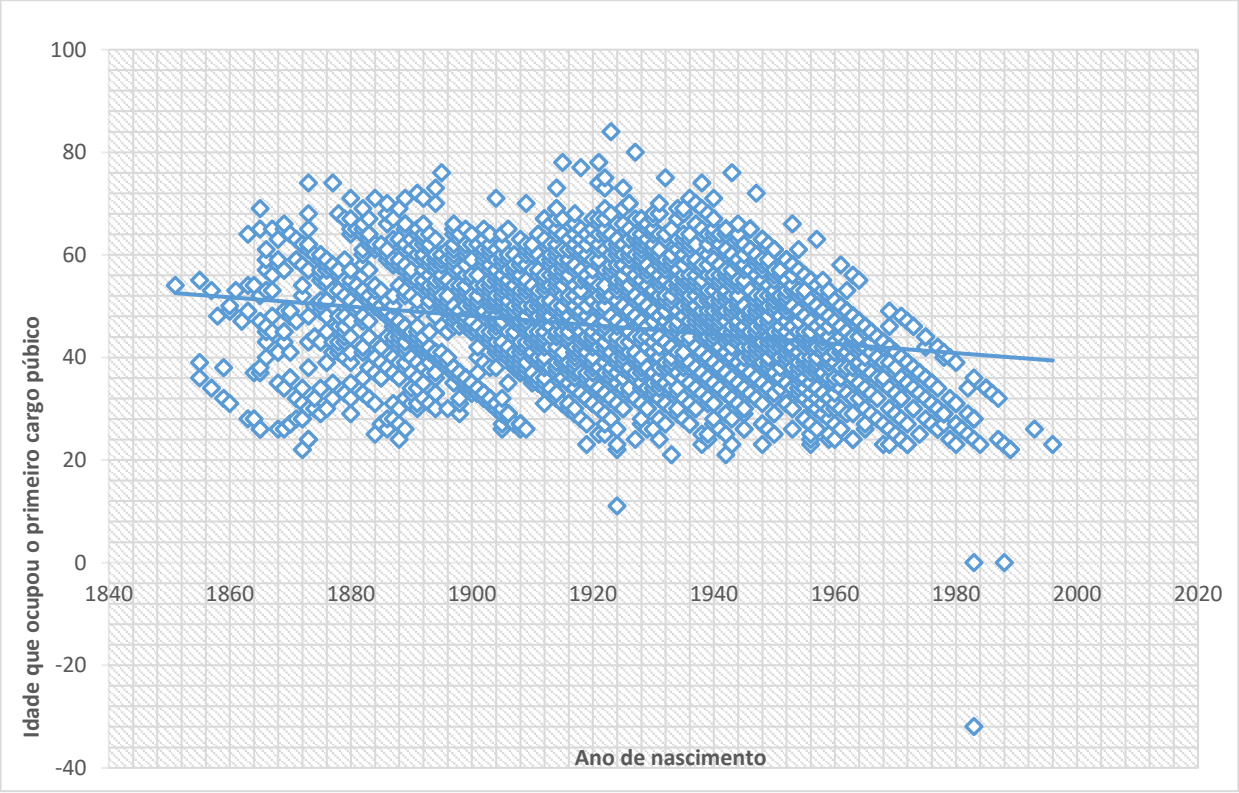

Figura 11 - Distribuição temporal dos indivíduos de acordo com a idade de início de carreira pública

Algo curioso que surgiu neste gráfico foi a presença de alguns *outliers* que claramente são erros que passaram desapercebidos no DHBB: um indivíduo está na posição próxima aos 10 anos de idade, dois na linha 0 e um no negativo. Ao verficar

os verbetes destes biografados, constatamos que de fato seus autores digitaram incorretamente a informação de ano de nascimento, já que, para além das idades negativas, dificilmente uma criança de 11 anos seria eleita deputada federal ou um jovem de 21 anos ocuparia nesta idade o cargo de ministro da Marinha. Estas incongruências servem como pistas para a identificação de eventuais erros nas informações contidas nos verbetes, o que possibilita seu conserto imediato.

A tabela abaixo traz as mesmas informações anteriores, mas restritas aos ocupantes dos cargos da esfera Executiva com distinção entre homens e mulheres:

	Nascidos em:	Total de indivíduos da amostra		Média de idade ao ocupar o primeiro cargo público (*qualquer* cargo público)
		Homens	Mulheres	
Presidentes da República	Antes de 1900	13	-	44,30
	1901 - 1920	8	-	43,75
	1921 - 1940	2	-	39
	1941 - 1960	2	-	38
	1961 – 1980	-	-	-
	1981 - 2000	-	-	-
Ministros	Antes de 1900	218	-	50,20
	1901 - 1920	245	1	49,31
	1921 - 1940	194	2	49,18
	1941 - 1960	159	20	46,51
	1961 – 1980	23	2	42,24
	1981 - 2000	1	-	34,0
Senadores	Antes de 1900	162	-	47,67
	1901 - 1920	200	-	44,46
	1921 - 1940	183	5	46,54
	1941 - 1960	139	12	42,74
	1961 – 1980	16	6	39,63
	1981 - 2000	-	-	-
Deputados Federais	Antes de 1900	536	2	46,56
	1901 - 1920	951	1	44,29
	1921 - 1940	1088	24	45,62
	1941 - 1960	1139	93	43,90
	1961 – 1980	269	32	37,73
	1981 - 2000	21	4	27,61

Tabela 30 - Média de idade do início de carreira dos políticos separados por gênero

A representação feminina nos principais papéis dessa elite política ainda é bastante tímida se comparada ao universo masculino: nesta versão do DHBB, apenas 204 desses papéis foram exercidos pelas mulheres contra 5.569 exercidos pelos homens, ou seja, exatos 3,5% se calculados estatisticamente. Entretanto, essa participação vem crescendo nas últimas décadas, sendo maior a presença daquelas que nasceram entre as décadas de 1940-1960, muitas provavelmente ainda no exercício de seus mandatos hoje. Elas se inserem na geração de políticos que começaram suas carreiras públicas na faixa dos 43 a 45 anos de idade. No caso do papel de deputados federais, onde a presença feminina é maior que nos outros cargos, percebemos que a cada geração a distância proporcional entre a quantidade de homens e mulheres torna-se mais estreita: se na geração dos nascidos entre 1901-1920 era de 951 homens x 1 mulher, nas gerações mais recentes essa proporção cai para 269 homens x 32 mulheres, ou ainda, 21 homens x 4 mulheres.

6.2
Qual a formação acadêmica dos políticos?

Foram compiladas 48 formações acadêmicas presentes nas sentenças recuperadas, conforme a tabela 12 mostrou. Muitas formações têm uma mesma raiz, variando apenas nas suas especializações ou na forma como é referenciada nos textos. A fim de melhorar o resultado das frequências criamos uma lista de correspondências e aplicamos aos trechos recuperados. Dessa forma, "ciências jurídicas" e "advogado", por exemplo, são incluídos em "direito"; "sociologia" em "ciências-sociais"; "cardiologista" em "medicina", e assim por diante. No final obtivemos uma visão mais enxuta das áreas percorridas pelos políticos. Abaixo, elencamos as 10 mais frequentes no DHBB[51].

Área de formação	Ocorrências
Direito	3851
Formação-militar	1082
Engenharia	912
Medicina	839
Economia	633
Administração	613
Filosofia	289

[51] A tabela completa com todas as ocorrências encontra-se no Anexo 2

Contabilidade	192
Letras	190
Ciências-sociais	147

Tabela 31 - Áreas de formação mais frequentes entre os biografados, em todo o período

Direito aparece em primeiro lugar, contabilizando mais que o triplo da segunda posição, que é a formação militar.

A seguir observamos a distribuição pelas seis gerações de políticos que discriminamos no exercício anterior, desconsiderando menções repetidas das áreas nos trechos (ou seja, mesmo que várias formações do tipo militar apareçam em um único trecho, apenas uma será contabilizada).

Área	TODOS	Verbetes sem data	Antes 1900 (geração 1)	1901-1920 (geração 2)	1921-1940 (geração 3)	1941-1960 (geração 4)	1961-1980 (geração 5)	1981-2000 (geração 6)
Direito	2.914	59	618	711	829	559	133	5
Militar	854	18	316	326	162	27	4	1
Engenharia	733	16	157	135	189	206	27	3
Medicina	672	19	157	163	122	195	15	1
Administração	516	11	7	35	132	256	71	4
Economia	517	11	10	73	132	258	32	1
Filosofia	267	7	26	63	97	64	10	-
Contabilidade	186	7	4	42	77	52	4	-
Letras	179	2	30	35	46	50	16	-
C. Sociais	137	2	9	17	40	55	13	1

Tabela 32- Distribuição das formações mais frequentes, por geração

A formação em direito aparece preponderante em todas as gerações de políticos, seguido da formação militar, com pouco menos de um terço. Os diplomas em engenharia, medicina, administração e economia vêm em seguida, sendo estes dois últimos praticamente empatados.

Em um artigo publicado em 2000, Fabiano Santos levantou dados biográficos dos deputados federais empossados no período pré e pós regime ditatorial, que apontaram para uma diminuição do número de advogados na Câmara a partir da redemocratização: durante a República de 1946, 57% dos deputados eram formados em Direito, este número subiu para 61% no período militar e caiu para menos de 40% na eleição de 1990 (Santos, 2000: 97). Estes índices são condizentes com a tabela acima, supondo que os primeiros desses deputados se localizam entre as gerações 1 e 2, os segundos nas gerações 2 e 3, e os últimos, nas

gerações 4 e 5. Simultaneamente à redução dos bacharéis em Direito, o autor também verificou um aumento da participação de outras formações, reflexo do crescimento da tecnocracia nos bastidores do poder, com a Engenharia passando de 6,6% no período anterior à ditadura para 8,2% com a redemocratização, Economia de 1,6% para 7,5%, e Medicina de 8,2% para 14,8% no mesmo período.

Na tabela percebemos que a transformação mais sentida foi o declínio na formação militar da segunda para a terceira geração, passando quase que pela metade. De fato, a menor incidência de políticos com formação militar a partir da terceira geração sugere que o treinamento civil estava substituindo o militar enquanto caminho mais adequado para atingir cargos políticos importantes.

Um estudo interessante que mostra as implicações das formações acadêmicas predominantes dos parlamentares para a política governamental é realizado por Neiva e Izumi (2012), que descrevem detalhadamente o perfil acadêmico dos senadores brasileiros no período de 1987 a 2006, trazendo dados de acordo com a distribuição regional, por partidos, por média de idade e por comissões temáticas. Um dos achados mais interessantes, segundo os autores, diz respeito ao predomínio dos economistas no PT e no PSDB. Enquanto mais de 20% dos membros das duas legendas estão nessa categoria, eles não passam de 8,6% nos outros partidos. No caso do PSDB, essa supremacia é dividida com os formados em Direito, o que permite ao partido ter uma atuação privilegiada nos assuntos econômicos; já no PT, a preponderância dos economistas é dividida com os profissionais da área de Humanidades, mais preocupados com os problemas sociais. Estes estudos mostram a importância do especialista do domínio na interpretação de dados quantitativos.

A tabela abaixo apresenta a distribuição das áreas pelos cargos selecionados como amostra.

	Distribuição das áreas mais frequentes pela nossa amostra de cargos (um biografado pode ocupar mais de um cargo e trazer mais de uma formação)				
	TODOS	Apenas presidentes da República	Apenas ministros	Apenas senadores	Apenas deputados federais
	6.745 (100%)	**26 (100%)**	**894 (100%)**	**742 (100%)**	**4.314 (100%)**

Direito	2.914	13	395	312	1783
Militar	854	6	171	32	125
Engenharia	733	1	120	72	358
Medicina	672	2	44	95	424
Administração	516	1	76	43	330
Economia	517	2	90	54	259
Filosofia	267	1	30	27	157
Contabilidade	186	-	13	27	152
Letras	179	3	22	26	117
C. Sociais	137	1	28	14	80

Tabela 33 - Distribuição das áreas de formação mais frequentes por cargos ocupados

No caso dos presidentes, onde a checagem é facilitada pelo número menor de indivíduos, apenas três casos se mostraram controversos: um, de José Linhares, em "matriculando-se na Escola de Medicina, que entretanto *abandonaria* no segundo ano", a identificação foi correta, porém incompleta dada a segunda proposição, um problema ainda difícil de resolver na extração automática. O outro caso, de Fernando Henrique Cardoso em "doutor em ciências sociais em 1961 pela Faculdade de Filosofia, Ciências e Letras da USP", colocou o presidente em três formações - ciências sociais, filosofia e letras - quando apenas uma, a primeira, é correta.

O problema em criar uma regra que ignore os nomes das faculdades, é que muitas vezes a única pista para descobrir a formação do indivíduo é justamente esta: "ingressou na Faculdade de Direito", "matriculou-se na Escola de Medicina". Mas uma ideia de implementação futura é fazer isto apenas para instituições que deem margem a tais dubiedades. Algo, aliás, muito semelhante ao que acontece no terceiro caso, em que "ingressou na Academia Maranhense de Letras" e "ingressar na Academia Brasileira de Letras" - presentes na biografia de José Sarney e Juscelino Kubitschek respectivamente -, levam ao erro de supor uma formação na área de Letras. Portanto, por causa do nome destas instituições, as formações em Letras e Filosofia estão inflacionadas nos verbetes.

6.3
O que dizer sobre os vínculos familiares na política?

Quem são os políticos que detêm vínculos familiares com outros políticos? Que vínculos são esses?

É impossível entender o Brasil e suas relações políticas sem compreender o papel das grandes famílias. Os elos de parentesco na política podem remontar ao período colonial, formando dinastias que se traduzem em fenômeno social empreendido como estratégia de manutenção dos espaços do poder, empurrando filhos e parentes às Câmaras e ao Senado (Schoenster, 2014; Medeiros, 2016; Oliveira et al, 2017). Esse fenômeno aumentou ou diminuiu ultimamente? As relações familiares têm variação quanto ao período do exercício no poder? Ou seja, varia entre as gerações de políticos ou períodos da História nacional? Quantas mulheres que estão na política têm laços de parentesco com outro político?

Segundo o cientista político Ricardo Costa Oliveira (Oliveira, 2018) que estuda a presença das famílias no poder, em 2018 cerca de 62% da Câmara é formada por deputados originários de famílias políticas, enquanto no Senado esse número sobe para mais de 70%. Ou seja, praticamente dois terços do Congresso brasileiro está tomado por algumas famílias.

Grande parte destas informações estão diluídas nos verbetes do DHBB, e identificá-las para fins de extração automática se mostrou um grande desafio. Começamos por adicionar anotações semânticas relativas a relações familiares e melhoramos a identificação dos nomes dos biografados a partir de regras de correspondência. Os resultados foram bastante bons, mas ainda é só o começo.

Abaixo reproduzimos alguns dados obtidos a partir das expressões utilizadas no exercício de extração de relações familiares no corpus. Dizem respeito apenas às relações que são realizadas entre biografados que possuem identificação no DHBB (e, portanto, são obrigatoriamente personagens públicos ou políticos). Foram recuperadas 1.190 ocorrências e sistematizadas no R juntamente com outros metadados.

dhbb_id	verbete	sexo	cargos	trechos	nasc
924	Elcival Ramos Caiado	m	@ dep. fed. GO 1975-1979	Seu irmão , Emival Ramos Caiado	1923
4104	Augusto Amaral Peixoto Júnior	m	@ militar @ rev. 1924 @ rev. 1930 @ const. 1934 @ dep. ...	Seu irmão , Ernâni Amaral Peixoto	1901
2305	Orlando Geisel	m	@ militar @ ch. Depto. Ger. Pess. Ex. 1965-1966 @ comte. I...	Seu irmão , Ernesto Geisel	1905
2837	Afonso Augusto de Albuquerque Lima	m	@ militar @ rev. 1930 @ min. Interior 1967-1969 @ ch. D...	Seu irmão , Estênio Caio de Albuquerque Lima	1909
4616	José Diogo Brochado da Rocha	m	@ militar @ rev. 1922 @ const. 1946 @ dep. fed. RS 1946...	Seu irmão , Francisco de Paula Brochado da Rocha	1904
5457	Benjamim Dornelles Vargas	m	@ dir. ger. DFSP 1945	Seu irmão , Getúlio Dornelles Vargas	1897
2076	Teodomiro Porto da Fonseca	m	@ rev. 1930 @ const. 1946 @ dep. fed. RS 1946-1951	Seu irmão , Gregório Porto da Fonseca	1879
3100	José Machado Sobrinho	m	@ dep. fed. MG 1971-1979 @ dep. fed. MG 1982 @ dep. f...	Seu irmão , Guilherme Machado	1932
4847	Ademar Santillo	m	@ dep. fed. GO 1975-1987	seu irmão , Henrique Santillo	1939

Figura 12 - Sistematização das relações familiares com outros metadados

Cabe lembrar que um biografado pode estabelecer mais de um vínculo familiar com outras pessoas, e isto está incluído nessas 1.190 menções. Se contabilizarmos apenas o número de políticos distintos que possui ao menos uma relação familiar com outro personagem, esse total cai para 858 indivíduos, o que representa 12,8% do total de biografados do DHBB.

Na tabela 26 sistematizamos estas informações, separando os totais por geração e cargo, indicando ainda o total de verbetes por categoria para obter uma estimativa da proporção:

Biografados com vínculo /total de biografados	Nascidos em:	Identificação de pelo menos um vínculo familiar com outro biografado /total de biografados			
		Homens	Mulheres	TODOS	%
TODOS 858 /6.717 **12,8 %**	SEM DATA	**22** /298	**1** /7	**23** /305	**7,5 %**
	Antes de 1900	**227** /1.334	**2** /6	**229** /1.340	**17 %**
	1901 - 1920	**222** /544	**2** /7	**229** /551	**41,5 %**
	1921 - 1940	**167** /1.552	**5** /30	**172** /1.582	**10,9 %**
	1941 - 1960	**141** /1.442	**17** /120	**158** /1.562	**9,9 %**
	1961 – 1980	**39**/ 316	**5** /39	**44** /355	**12,4 %**
	1981 - 2000	**2** /8	**1** /4	**3** /12	**25 %**
Presidentes da República 13 /26 **50 %**	SEM DATA	**0** /2	-	**0** /2	**0 %**
	Antes de 1900	**6** /12	-	**6** /12	**50 %**
	1901 - 1920	**5** /8	-	**5** /8	**62,5 %**
	1921 - 1940	**2** /2	-	**2** /2	**100 %**
	1941 - 1960	**0** /2	-	**0** /2	**0 %**
	1961 – 1980	-	-	-	**0 %**
	1981 - 2000	-	-	-	**0 %**
Ministros 168 /894 **18,8 %**	SEM DATA	**4** /20	-	**4** /20	**20 %**
	Antes de 1900	**58** /231	-	**58** /231	**25,1 %**
	1901 - 1920	**50** /232	**0**/ 1	**50** /233	**21,4 %**
	1921 - 1940	**29** /204	**0** /2	**29** /202	**14,3 %**
	1941 - 1960	**22** /161	**3** /21	**25** /182	**13,7 %**
	1961 – 1980	**2** /23	**0** /2	**2** /25	**8 %**
	1981 – 2000	**0** /1	-	**0** /1	**0 %**
Senadores 260 /742 **35 %**	SEM DATA	**5** /17	**0** /1	**5** /18	**27,7 %**
	Antes de 1900	**89** /169	-	**89** /169	**52,6 %**
	1901 – 1920	**72** /193	-	**72** /193	**37,3 %**
	1921 – 1940	**56** /183	**0** /5	**56** /188	**29,8 %**
	1941 – 1960	**31** /139	**4** /12	**35** /151	**23,2 %**
	1961 – 1980	**3** /17	**0** /6	**3** /23	**13,0 %**
	1981 - 2000	-	-	-	-
Deputados Federais 781 /4.314 **18,1 %**	SEM DATA	**13** /152	**0** /4	**13** /156	**8,3 %**
	Antes de 1900	**182** /580	**0** /2	**182** /582	**31,3 %**
	1901 - 1920	**175** /907	**0** /1	**175** /908	**19,3 %**
	1921 - 1940	**186** /1.088	**5** /24	**191** /1.112	**17,2 %**
	1941 - 1960	**142** /1.139	**20** /93	**162** /1.232	**13,2 %**
	1961 – 1980	**47** /269	**7** /32	**54** /301	**17,9 %**
	1981 - 2000	**2** /19	**2** /4	**4** /23	**17,4 %**

Tabela 34 - Média de idade do início de carreira dos políticos separados por gênero

Ainda que com esta amostra não seja possível determinar quão entranhada está a política brasileira no que tange às dinastias familiares, já que a lógica de domínio pelo parentesco também se dá em outras esferas de poder que o DHBB muitas vezes não abrange (como os executivos e legislativos estaduais e municipais), é possível perceber que se trata de um fenômeno que se mantêm ao longo do tempo, em índices bastante significativos.

Presidentes e senadores são os políticos que mais aparecem com vínculos familiares, 50% e 35% respectivamente, sendo mais percebido nas primeiras

gerações (nesta versão do DHBB). Ministros e deputados seguem na casa dos 18%, mantendo a média ao longo das gerações. No cômputo geral, foram contabilizadas 33 mulheres, de um total de 213, identificadas com algum vínculo com outro personagem político, o que corresponde a 15,5%. É ligeiramente mais alto que o índice de 14,9% obtido pelos homens, que totalizaram 820 de um total de 5.494.

Infelizmente a maior parte dos políticos eleitos em 2018 ainda não foram incluídos nesta versão, e, portanto, não há como comparar os números obtidos nesta extração com os que Oliveira identificou após as eleições para a Câmara e o Senado, de 62% e 70% respectivamente, de indivíduos originários de famílias políticas (Oliveira, 2018).

7
Considerações finais

Esta tese buscou investigar a possibilidade de extrair informação útil, diversificada e de alta qualidade a partir de um corpus de estilo enciclopédico. Conduziu estudos sobre anotações em corpora e abordagens para extração automática de informações, e procurou identificar os desafios e as contribuições teóricas e metodológicas que melhor endereçam a tarefa. Descreveu uma metodologia baseada no uso de padrões textuais, onde regras de extração são desenvolvidas manualmente apoiadas principalmente nos aspectos léxico-sintáticos e anotações semânticas atribuídas ao material. Testou a abordagem proposta em um corpus de estilo enciclopédico no domínio da história contemporânea do Brasil, em três exercícios de extração pontuais.

De um modo geral, os resultados apresentaram alta precisão na qualidade das informações localizadas. Dentre as dificuldades percebidas, encontrar o equilíbrio entre o número suficiente de padrões *versus* a abrangência esperada e avaliar e tratar de forma sistemática os resultados obtidos, figuram como principais. A primeira tem a ver com a quantidade – e por consequência, o trabalho manual demandado – de expressões que deverão ser criadas e aplicadas no corpus, o que não é algo fácil de prever pela própria expressividade natural da língua. E a segunda relaciona-se aos falsos positivos, onde o padrão foi corretamente identificado no texto, porém a informação recuperada não se mostrou verdadeira: por exemplo, quando determinado evento foi atribuído ao biografado, mas não deveria (caso das anáforas ou do sujeito nulo). O desenvolvimento de novos recursos e ferramentas de PLN que sejam capazes de melhorar e/ou facilitar a identificação das estruturas de informação de interesse nos textos, combinado à adoção de outras técnicas de extração pode vir a otimizar o processo e fazer com que a qualidade da informação recuperada atinja níveis maiores de confiança.

No capítulo introdutório, foram colocados exemplos de informações que seriam desejáveis extrair dos verbetes, partindo do interesse real de alguns pesquisadores. Estes exemplos, na forma de perguntas, são retomados abaixo a título de avaliação do que seria possível responder a partir da extração de informação.

1. Quais os políticos que nasceram antes da década de 1960, tiveram formação militar e ocuparam algum cargo no Executivo?
2. Como se caracteriza a formação superior dos quadros políticos ao longo das gerações?
3. Qual a idade dos ministros do Supremo Tribunal Federal ao serem nomeados?
4. Qual o perfil partidário dos ministros do Poder Executivo republicano?
5. Quem são os políticos que detêm vínculos familiares com outros políticos? Que vínculos são esses?

As perguntas de 1 a 3 podem se beneficiar do primeiro e segundo exercícios realizados na seção 5.4.3, pois dizem respeito às mesmas informações compreendidas ali: data de nascimento e formação acadêmica, além dos cargos que são fornecidos pelos metadados dos verbetes. A pergunta 4 é um pouco mais complexa, pois tem a ver com o que se entende por "perfil partidário". Para além da composição social dominante nos partidos, existe o fato de que os parlamentares mudam de legenda ao longo da sua trajetória, e os próprios partidos mudam de nome como uma estratégia para se aproximarem dos eleitores, e nem sempre tais informações são construídas de forma explícita nos textos. Além disso, partidos são agrupados de acordo com a sua orientação ideológica (direita, centro, esquerda, etc.), segundo critérios utilizados por pesquisadores, pela mídia e pelos próprios políticos, constituindo-se como uma informação fluida.

Sobre a pergunta 5, explorada em parte no último exercício de aplicação da metodologia proposta, foi mostrado que é possível recuperar bastante informação com resultados promissores. Para isso foi fundamental a criação do léxico de família para a localização dessas relações no corpus, que mostrou que o enriquecimento semântico eleva o potencial da abordagem e indicando que há espaço para novas explorações.

Como trabalhos futuros, pretende-se incluir mais camadas de anotações semânticas no corpus, à semelhança do que foi feito com relações familiares, abrangendo, por exemplo, vínculos institucionais, eventos relacionados à

corrupção, condenações, redes de influência, localizações georreferenciadas etc. Também serão feitos novos processos de *grounding* para a desambiguação de nomes próprios de lugares, eventos, partidos políticos e papéis.

Pretende-se ainda aprofundar as análises dos temas propostos com estudos e referenciais teóricos encontrados na história e ciência política, a fim de confrontar ou promover melhor compreensão dos resultados obtidos. Outros cruzamentos de informações poderão ser testados, como por exemplo, distribuição das formações acadêmicas por partido e região, distribuição dos vínculos familiares por legislatura, distribuição da idade dos políticos no exercício de determinado(s) cargo(s), etc.

Exploração com outros esquemas de anotação, outros analisadores e outras abordagens de extração (combinando padrões textuais e aprendizado de máquina) estão na agenda de trabalho a título de comparação de resultados, metodologia e expansão do corpus. A ideia é retomar os desafios elencados na seção 4.3.2, tais como ambiguidade, vagueza, anaforismo e sujeito oculto, e buscar as melhores soluções para contorná-los. Por fim, as anotações e os resultados apresentados poderão ser utilizados e trazer avanços em aplicações diversas, como sistemas de recuperação da informação, perguntas e respostas, análise de redes e georreferenciamento.

Embora a metodologia em si não seja inédita, a ideia do experimento foi, ademais, explorar questões que têm sido endereçadas nos debates da chamada história digital, conforme vimos na seção 2.1: o quanto o fazer história tem sido modificado com o uso de ferramentas digitais, os desafios que precisam ser enfrentados e as oportunidades que se abrem neste cenário potencialmente inovador. Ao nosso ver, estas novas ferramentas podem sim suscitar a ampliação da pesquisa acadêmica nas ciências sociais e humanas sob diversos aspectos, tanto em termos de renovação de métodos quanto de produção de conhecimento. Mas a disposição para experimentar novos caminhos, especialmente no campo das humanidades digitais, deve vir acompanhada desta compreensão: não se trata de renunciar ou se opor às abordagens tradicionais de investigação, mas de perceber que a complementaridade entre os métodos computacionais e os mais artesanais ou

humanos tem potencial tanto para ajudar a responder questões antigas quanto para produzir novas questões.

As áreas das ciências humanas – e a história em particular –, sempre legaram aos registros textuais grande parte da sua razão de ser e de seu modo de fazer. No escrutínio das leituras e na produção da escrita, o saber é retido, somado e construído com a ajuda inestimável das fontes e dos recursos disponíveis. Aprende-se que o ofício é regido por métodos e técnicas condizentes a cada época, afinal, "cada sociedade se pensa historicamente com os instrumentos que lhe são próprios" (Certeau, 1988:28).

Por fim, as dificuldades existem e os desafios são muitos, mas as possibilidades abertas por cenários de inovação levam a um contínuo e merecido esforço para tentar superá-los.

8
Referências bibliográficas

ABREU, A. A.; BELOCH, I.; LAMARAO, S. T. N.; LATTMAN-WELTMAN, F.; PAULA, C. J. (orgs). Dicionário Histórico-Biográfico Brasileiro Pós-1930, Rio de Janeiro: FGV, 2010.

AGT-RICKAUER, H. Supporting Domain Modeling with Automated Knowledge Acquisition and Modeling Recommendations (Thesis). Elektrotechnik und Informatik der Technischen Universitat Berlin, 2019.

ALVES, D. "História e Humanidades Digitais: conexões para um novo tempo" (Entrevista). Entrevista concedida a Bruno Leal Pastor de Carvalho. In: Café História – história feita com cliques. Disponível em: https://www.cafehistoria.com.br/historia-e-humanidades-digitais. Publicado em: 17 Jul 2017. Acesso em: 25/05/2020.

ALMEIDA, Maurício B; SOUZA, Renato Rocha. Avaliação do espectro semântico de instrumentos para organização da informação. Encontros Bibli, v. 16, p. 25-50, 2011.

ANTHONY, L. A critical look at software tools in corpus linguistics. Linguistic Research 30(2), 2013.

ARAÚJO, N. "Vista de longe, a literatura é o que desaparece. (Acerca de um fracasso programático em Franco Moretti)". In: Andréa Sirihal Werkema, Marcus Vinícius Nogueira Soares, Nabil Araújo (orgs). Variações sobre o romance. Rio de Janeiro: Edições Makunaima, 2016.

ARCHER, D. "Corpus annotation: a welcome addition or an interpretation too far?" In: J. Tyrkkö, M. Kipiö, T. Nevalainen and M. Rissanen (eds.). Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources. Studies in Variation, Contacts and Change in English eSeries. 2012. Disponível em: http://www.helsinki.fi/varieng/journal/volumes/10/archer/ . Acesso em: 12/11/2020.

BARTHES, R. "Introdução à análise Estrutural da narrativa". In: BARTHES, R. Análise estrutural da narrativa, 19-60, 1976.

BICK, E. The parsing system palavras: Automatic grammatical analysis of Portuguese in a constraint grammar famework. Aarhus Universitetsforlag, 2000.

BICK, E. "Functional Aspects on Portuguese NER". In Diana Santos & Nuno Cardoso (eds.), Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, pp. 145-155, 2007.

BONELLI, E. T. “Theoretical overview of the evolution of corpus linguistics”. In: McCarthy, Michael & O'Keeffe Anne. The Routledge Handbook of Corpus Linguistics. UK, Taylor & Francis e-Library, 2010.

BONFIGLIOLI, R.; NANNI, F. “From Close to Distant and Back: How to Read with the Help of Machines”. 3rd International Conference on History and Philosophy of Computing (HaPoC), Pisa, Italy, 2015.

BRASIL, E.; NASCIMENTO, L. F. "História digital: reflexões a partir da Hemeroteca Digital Brasileira e do uso de CAQDAS na reelaboração da pesquisa histórica." Revista Estudos Históricos, v. 33, n. 69, 2020.

BURKE, P. A escola dos Annales (1929-1989). Unesp, 1997.

CAMBRIA, E.; PORIA, S.; BISIO, F.; BAJPAI, R.; CHATURVEDI, I.” The CLSA Model: A Novel Framework for Concept-Level Sentiment Analysis”. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 3-22). Springer, 2015.

CARDOSO, N.; SANTOS, D. “Directivas para a identificação e classificação semântica na colecção dourada do HAREM”. In Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área, 2007, pp. 211–238.

CARR, E. H. Que é história?. Conferências George Macaulay Trevelyan, Universidade de Cambridge. Paz e Terra, 1978.

CARVALHO, P.; OLIVEIRA, H. G.; SANTOS, D.; FREITAS, C.; MOTA, C. “Segundo HAREM: Modelo geral, novidades e avaliação”. In: Mota, Cristina & Santos, Diana (orgs). Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: o Segundo HAREM. 2008.

CASTRO, C.; Higuchi, S.; Monnerat, S. A obra de Gilberto Velho: uma leitura distante para observar o familiar. CPDOC, Rio de Janeiro, 2021.

CERTEAU, M. “A operação histórica”. In: NORA, Pierre & LE GOFF, Jacques (Org.). História: novos problemas. Rio de Janeiro: F. Alves, 1988.

CHINCHOR, N. MUC-7 Named Entity Task Definition. Version 3.5, 1997. Disponível online: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html, acesso em 5 de março de 2017.

CHITICARIU, L.; YUNYAO, L.; REISS, F. "Rule-based information extraction is dead! long live rule-based information extraction systems!". In Proceedings of the 2013 conference on empirical methods in natural language processing, 2013.

CHOWDHURY, G. G. Natural language processing. Annual review of information science and technology, 37(1), 51-89, 2003.

COATES-STEPHEN, S. The analysis and acquisition of proper names for robust text understanding. Doctoral thesis, City University London, 1992. Disponível em http://openaccess.city.ac.uk/8015/. Acesso em 20/07/2019.

COHEN, D.J., FRISCH, M., GALLAGHER, P., MINTZ, S., SWORD, K., TAYLOR, A.M., THOMAS, W.G.; TURKEL, W.J.. "Interchange: The promise of digital history'. The Journal of American History, 95(2), 2008.

CONNIFF, M. "The national elite". In: CONNIFF, M.; McCANN, F. (eds). Modern Brazil: elites and masses in historical perspective. U of Nebraska Press, 1991.

CONNIFF, M. "O DHBB e os brasilianistas". In: FGV, E. (ed.) CPDOC 30 Anos. Editora FGV/CPDOC, Rio de Janeiro, 2003.

CONNIFF, M. A elite nacional. Por outra história das elites. Rio de Janeiro: FGV, 2006.

COSTA, L., SANTOS, D., & ROCHA, P. A. "Estudando o português tal como é usado: o serviço AC/DC". In The 7 th Brazilian Symposium in Information and Human Language Technology (STIL 2009). São Carlos, 2009.

CRYMBLE, A. "Historians are becoming computer science customers postscript", Digital History Seminar, 2015. Disponível em: http://ihrdighist.blogs.sas.ac.uk/2015/06/24/historians-are-becoming-computer-science-customers-postscript/. Acesso em 25/12/2020.

CUNNINGHAM, H. "Information Extraction, Automatic". In Encyclopedia of Language and Linguistics, 2nd Edition, Elsevier, 2006.

CURRAN, J.; CLARK, S. J. "Language independent NER using a maximum entropy tagger". In Proc. CoNLL-2003, 2003

DOBSON, J. E. "Can an Algorithm be Disturbed? Machine Learning, Intrinsic Criticism, and the Digital Humanities". In College Literature. 42 (4): 543–564, 2015.

EVERT, S. e The CWB Development Team. "CQP Query Language Tutorial", IMS Stuttgart, Maio 2019. Disponível em http://cwb.sourceforge.net/files/CQP_Tutorial.pdf. Acesso em 18/06/2019.

FAIRCLOUGH, Norman. Discurso e mudança social. Brasília: Editora UnB, 2008.

FLORIAN, R., Ittycheriah, A., Jing, H., & Z"hang, T. "Named entity recognition through classifier combination". In Proceedings of conll-2003 (p. 168-171). Edmonton, Canada, 2003.

FORTES, A.; ALVIM, L. G. M. "Evidências, códigos e classificações: o ofício do historiador e o mundo digital." In Esboços: histórias em contextos globais v. 27, n.45, 2020.

FREITAS, C. Elaboração automática de ontologias de domínio: discussão e resultados. Tese de Doutorado. Pontifícia Universidade Católica do Rio de Janeiro, 2007

FREITAS, C., SANTOS, D. e GONÇALVES, A. Perguntas já respondidas sobre o AC/DC: desde como começar até o uso complexo de funcionalidades poderosas. 2011. Versão 2.1 disponível em: https://www.linguateca.pt/acesso/PJR_ACDC_Tudo.pdf.

FREITAS, C.; SANTOS, D.; OLIVEIRA, H. G.; CARVALHO, P.; MOTA, C. "Relações semânticas do ReRelEM: além das entidades no Segundo HAREM" in MOTA, Cristina & SANTOS, Diana (eds.). Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

FREITAS, C.; CARVALHO, P.; OLIVEIRA, H. G.; MOTA, C.; SANTOS, D. "Second HAREM: advancing the state of the art of named entity recognition in Portuguese". In Nicoletta Calzolari et al. (eds.), Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010). European Language Resources Association, pp. 3630-3637. Valletta, 2010.

FREITAS, C. Esqueleto: investigação sobre o léxico do corpo para a inclusão de informação semântica em corpora da língua portuguesa. Detalhamento do plano de pesquisa., s/d.

FREITAS, C. Corpus, "Linguística Computacional e as Humanidades Digitais". In: Leite, M. e Gabriel, C. T. (orgs). Linguagem, Discurso, Pesquisa e Educação. Rio de Janeiro, DP et ali, 2015.

FREITAS, C. Elaboração automática de ontologias de domínio: discussão e resultados. Tese de Doutorado. Pontifícia Universidade Católica do Rio de Janeiro, 2007

GARSIDE, R.; LEECH, G.; McENERY, A. (eds). Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman. 1997.

GIBBS, F.; OWENS, T. The hermeneutics of data and historical writing. Writing history in the digital age, v. 159, 2013.

GIOVANNETTI, E.; MARCHI, S.; MONTEMAGNI, S. Combining Statistical Techniques and Lexico-syntactic Patterns for Semantic Relations Extraction from Text. In: SWAP. 2008.

GIRJU, R. "Automatic detection of causal relations for question answering". In: Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering. Stroudsburg, PA, 2003.

GOLD, M. K. "Day of DH: Defining the Digital Humanities". In GOLD, M. K. (ed). Debates in the Digital Humanities. University of Minnesota Press, 2012.

Disponível em https://dhdebates.gc.cuny.edu/projects/debates-in-the-digital-humanities. Acesso em 26/06/2020.

GOLD, M. K.; KLEIN, L. Digital Humanities: The Expanded Field. In GOLD, M. K.; KLEIN, L. (orgs). Debates in the digital humanities 2016. University of Minnesota Press, 2016.

GOLD, M. K.; KLEIN, L. A. "DH That Matters". In GOLD, M. K.; KLEIN, L. (orgs). Debates in the digital humanities 2019. University of Minnesota Press, 2019.

GOLSHAN, P. N., DASHTI, H. R., AZIZI, S., & SAFARI, L. A Study of Recent Contributions on Information Extraction. arXiv preprint :1803.05667. 2018

GRACIOSO, L. S. & SALDANHA, G. Ciência da Informação e Filosofia da Linguagem: da pragmática informacional à web pragmática. 2010.

GRISHMAN, R. "Information Extraction". In IEEE Intelligent Systems. New York, pp. 8-15, May 2015.

GRUBER, J. Markdown language. 2014. Disponível em: http://daringfireball.net/projects/markdown/ Acesso em 23/05/2019.

HAMMOND, A. The double bind of validation: distant reading and the digital humanities "through of disillusionment". Literature Compass, v. 14, n. 8, p. e12402, 2017.

HARRIS, Zellig S. Distributional Structure, WORD, 10:2-3, 146-162, 1954.

HEARST, M. "Automatic acquisition of hyponyms from large text corpora". In Proceedings of the 14th conference on Computational linguistics - Volume 2, COLING '92, pages 539–545, Stroudsburg, PA, USA, 1992.

HEARST, M. "Automated discovery of WordNet relations", in Fellbaum, Christiane, ed., WordNet: An Electronic Lexical Database, MIT Press, May 1998.

HIGUCHI, S.; Santos, D.; Freitas, C. & Rademaker, A. "Distant reading brazilian politics". In CEUR Workshop Proceedings. ISSN 1613-0073. 2364, s 191- 200, 2019

HIRST, G. "The future of text-meaning in computational linguistics". In Lecture Notes in Computer Science, Vol. 5246. Springer-Verlag Berlin Heidelberg, 2008.

HIRST, G. "Limitations of the Philosophy of Language Understanding Implicit in Computational Linguistics". In: Proceedings, VIIth European Conference on Computing and Philosophy, Barcelona, 2009.

HOCKEY, S. The history of humanities computing. Susan Schreibman, Ray Siemens, John Unsworth (eds). Oxford: Blackwell, 2004. (http://www.digitalhumanities.org/companion/)

IDE, N. & VERONIS, J. “Introduction to the special issue on word sense disambiguation: The state of the art”. Computational Linguistics 24:1-40, 1998.

JACOBS, P.; Zernik, U. “Acquiring lexical knowledge from text: A case study”. In Proceedings of AAAI88, 1988.

JANICKE, S.; FRANZINI, G.; CHEEMA, M. F.; SCHEUERMANN, G. "On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges." In EuroVis (STARs), pp. 83-103. 2015.

JOCKERS, M. Macroanalysis. Digital Methods and Literary History. Chicago: U of Illinois P, 2013.

JURAFSKY, D., MARTIN, J. H. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River: Pearson/Prentice Hall, 2009.

KILGARRIFF, Adam. I don't believe in word senses. Brighton, 2003.

KIRSCH, A. "Technology Is Taking Over English Departments". The New Republic. The New Republic, 2014. Disponível em https://newrepublic.com/article/117428/limits-digital-humanities-adam-kirsch. Acesso em 7/7/2017.

KUPIEC, J. “A robust linguistic approach for question answering using an on-line encyclopedia.” In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 181-190, Pittsburgh, 1993.

LACAN, J. (1998) Escritos, Rio de Janeiro, Jorge Zahar.

LADURIE, E. L. “O historiador e o computador”. In: NOVAIS, Fernando; SILVA, Rogério Forastieri da. Nova história em perspectiva. São Paulo: Cosac Naify, 2011. p. 207-210, v. 1.

LDC Linguistic Data Consortium. English Annotation Guidelines for Entities. ACE (Automatic Content Extraction). 2008. Disponível: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf. Acessado em 07/03/2019.

LEECH, G. “Introducing Corpus Annotation”. In: GARSIDE et al. Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman. 1997

LEECH, G., “Adding Linguistic Annotation”. In: WYNNE, M (editor). Developing Linguistic Corpora: a Guide to Good Practice, 2005. Disponível em: http://www.ahds.ac.uk/guides/linguistic-corpora/index.htm

LUCCHESI, A. "Por um debate sobre História e Historiografia Digital." Boletim Historiar 2, 2014.

MAKAROV, P. “Automated Acquisition of Patterns for Coding Political Event Data: Two Case Studies”. In Proceedings of Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 103–112. Santa Fe, New Mexico, USA, 2018.

McCARTHY, M.; O'KEEFFE, “A. Historical perspective: what are corpora and how have they evolved?”. In McCARTHY, M. & O'KEEFFE (eds). The Routledge Handbook of Corpus Linguistics. UK, Taylor & Francis e-Library, 2010

MANNING, C.; SCHUTZE, H. Foundations of Statistical Natural Language Processing. MIT Press, 1999.

MARCONDES, D. Textos básicos de linguagem: de Platão a Foucault. Rio de Janeiro: Zahar, 2009. E-book KINDLE.

MARCUSCHI, L. A. “Anáfora indireta: o barco textual e suas âncoras”. In Revista Letras, v. 56, 2001. Disponível em: https://revistas.ufpr.br/letras/article/view/18415/11987. Acesso em: 08/01/2021.

MARTINS, H. Três caminhos na filosofia da linguagem. in MUSSALIN, F e BENTES, A. Introdução a linguística – Fundamentos Epistemológicos, vol III, Cortez Editora, 2005.

MARTINS, F; FREITAS, C. “Sujeitos ocultos em verbetes biográficos: contornando dificuldades da extração automática de informações”. In XI Congresso Internacional da Abralin, 2019.

MARRERO, M; SÁNCHEZ-CUADRADO S.; LARA, J.M.; ANDREADAKIS, G. “Evaluation of Named Entity Extraction Systems”. Computer Engineering Department, University Carlos III of Madrid, 2009.

McENERY, T.; HARDIE, A. Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press. 2012.

MELLO, H. R.; SOUZA, R. R. "A linguagem da ciência: prospecção de dados baseados em corporas." Texto Livre: Linguagem e Tecnologia 7, no. 1 (2014): 158-166.

MIKHEEV, A., GROVER, C.; MOENS, M. “Description of the LTG system used for MUC-7”. In Proceedings of 1999 muc-7. University of Edinburgh, 1999.

MONTEIRO, J. M. A Política como Negócio de Família: os herdeiros e a força dos capitais no jogo político das elites na paraíba (1985-2015). Tese de doutorado. Campina Grande: Universidade Federal de Campina Grande, 2016.

MOTA, C. "Corte e costura no AC/DC: auxiliando a melhoria da anotação nos corpos". 14 de abril de 2014. http://www.linguateca.pt/acesso/corte-e-costura.pdf

MORETTI, F. “Conjectures on world literature”. New Left review, 2000.

MORETTI, F. Distant reading. Verso Books, 2013.

MORIN, E.; JACQUEMIN, C. “Automatic acquisition and expansion of hyperonym links”, in Computer and the Humanities, vol. 38 (4), 343-362l, 2004.

NADEAU, D., TURNEY, P. D., & MATWIN, S. “Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity”. In Conference of the Canadian society for computational studies of intelligence (pp. 266-277). Springer, Berlin, Heidelberg, 2006.

NAJAFABADI, M.; VILLANUSTRE, F.; KHOSHGOFTAAR, T.M. et al. “Deep learning applications and challenges in big data analytics”. Journal of Big Data 2, 1, 2015. Disponível em https://doi.org/10.1186/s40537-014-0007-7. Acessado em 03/11/2020.

NAKAMURA, J.; M. Nagao. “Extraction of semantic information from an ordinary english dictionary and its evaluation”. In Proceedings of the Twelfth International Conference on Computational Linguistics, pages 459{464, Budapest, 1998.

NEIVA, P.; IZUMI, M. Os" doutores" da federação: formação acadêmica dos senadores brasileiros e variáveis associadas. Revista de Sociologia e Política, 20(41), 171-192, 2012.

NOIRET, S. História Pública Digital │ Digital Public History. Liinc em Revista, v. 11, n. 1, 2015.

O'KEEFFE, A.; MCCARTHY, M. (Orgs.). The Routledge Handbook of Corpus Linguistics. USA: Routledge, 2012.

OLIVEIRA, R. C. et al. “Família, parentesco, instituições e poder no Brasil: retomada e atualização de uma agenda de pesquisa”. Revista Brasileira de Sociologia, v. 5, n. 11, p. 165-198, 2017. Disponível em https://dialnet.unirioja.es/descarga/articulo/6227086.pdf. Acessado em 20/05/2020.

OLIVEIRA, R. C. “Pesquisa sobre famílias na política escancara a perpetuação do poder”. APUFPR SSind, 2018. Disponível em: https://youtu.be/2OKHGU5yOBk . Acessado em 22/06/2019.

PAIVA, V.D.; OLIVEIRA, D.; HIGUCHI, S.; RADEMAKER, A.; MELO, G.D. “Exploratory information extraction from a historical dictionary”. In: IEEE 10th International Conference on e-Science (e-Science). vol. 2, pp. 11{18. IEEE (2014)

PENNYCOOK, A. “Os limites da linguística”. In: Lopes da Silva, F. e Rajagopalan, K. (eds.) A Lingüística que Nos Faz Falhar. São Paulo: Parábola, 2004, p. 39-43

PROVOST, F.; FAWCETT, T. Data Science for Business: what you need to know about data mining and data-analytic thinking. O'Reilly: California, 2013.

PUSTEJOVSKY, J.; STUBBS, A. Natural Language Annotation for Machine Learning. O’Reilly Media, 2012.

RADEMAKER, A.; OLIVEIRA, D.A.B.; de PAIVA, V.; HIGUCHI, S. “A linked open data architecture for the historical archives of the getulio vargas foundation”. International Journal on Digital Libraries 15(2-4), 153:167, 2015.

RADEMAKER, A.; CHALUB, F.; FREITAS, C. “Two Corpus Based Experiments with the Portuguese and English Wordnets”. In LDK Workshops (pp. 134-145), 2017.

RAMISCH, C. "A generic and open framework for multiword expressions treatment: from acquisition to applications." PhD diss., Université de Grenoble (France); Universidade Federal do Rio Grande do Sul, 2012.

RIBEIRO, C. J. S.; HIGUCHI, S.; FERLA, L. “Aproximações ao cenário das humanidades digitais no Brasil”. In Digital Humanities Quarterly, v. 14, p. 1-10, 2020.

ROMÃO, L. C S. Reconhecimento de entidades Mencionadas em Língua Portuguesa: Locais, Pessoas, Organizações e Acontecimentos. Instituto Superior Técnico, Universidade Técnica de Lisboa. Novembro 2007.

RORTY, Richard. Objetivismo, relativismo e verdade. Tradução Marco Antônio Casanova. Rio de Janeiro, Relume-Dumará (Escritos Filosóficos), 1997.

SAG, I.A.; BALDWIN, T.; BOND, F.; COPESTAKE, A. & FLICKINGER, D. “Multiword expressions: A pain in the neck for NLP”. In International conference on intelligent text processing and computational linguistics (pp. 1-15). Springer, Berlin, Heidelberg, 2002.

SAMPSON, G. Empirical Linguistics. London: Continuum, 2001

SANG, E.,” Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition”. In: Proceedings of CoNLL-2002, Taipei, Taiwan, 2002, pp. 155-158.

SANG, E.; DE MEULDER, F., “Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition”. In: Proceedings of CoNLL-2003, Edmonton, Canada, 2003, pp. 142-147.

SANTOS, F. “Deputados federais e instituições legislativas no Brasil: 1946-99”. In: SANTOS, F.; BOSCHI, R. & DINIZ, E. (orgs.). Elites políticas e econômicas no Brasil contemporâneo. São Paulo: Fundação Konrad Adenauer, 2000.

SANTOS, D.; MAMEDE, N.; BAPTISTA, J. “Extraction of family relations between entities”, Proceedings of the INForum, 2010.

SANTOS, Diana (ed.). Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa, IST Press, 2007.

SANTOS, D. "O modelo semântico usado no Primeiro HAREM". p. 43-57. In: SANTOS, Diana e CARDOSO, Nuno (eds). Reconhecimento de entidades

mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, 2007.

SANTOS, D. "Corporizando algumas questões". In Stella E. O. Tagnin & Oto Araújo Vale (orgs.), Avanços da Lingüística de Corpus no Brasil, Editora Humanitas/FFLCH/USP, São Paulo, 2008.

SANTOS, D. “Enquadramento e historial do Segundo HAREM”. In: Mota, Cristina & Santos, Diana (orgs). Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: o Segundo HAREM. 2008a.

SANTOS, D. “Linguatecas infrastructure for portuguese and how it allows the detailed study of language varieties”. In: J. B. Johannessen (ed.), Language Variation Infrastructure, Oslo Studies in Language 3(2), 2011.

SANTOS, D. “Podemos contar com as contas?”, in Sandra Aluísio & Stella Tagnin (eds.), New Language Technologies and Linguistic Research: A Two-way Road, Cambridge Scholars Publishing, 2014.

SANTOS, D. “Literature studies in Literateca: between digital humanities and corpus linguistics”. In Martin Doerr, Øyvind Eide, Oddrun Grønvik & Bjørghild Kjelsvik (eds.), Humanists and the digital toolbox: In honour of Christian-Emil Smith Ore. Novus forlag, Oslo, 2019, pp. 89-109.

SANTOS, D., & RANCHHOD, E. “Ambientes de processamento de corpora em português: Comparação entre dois sistemas”. In Irene Rodrigues; Paulo Quaresma (ed). Actas do IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR'99). Évora, 1999.

SANTOS, D., & BICK, E. “Providing Internet access to Portuguese corpora: the AC/DC project”. In Maria Gavrilidou; George Carayannis; Stella Markantonatou; Stelios Piperidis; Gregory Stainhauer (ed). Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000), Athens, 2000.

SANTOS, D., CARDOSO, N. "Breve introdução ao HAREM". p. 1-16. In: SANTOS, Diana e CARDOSO, Nuno (eds). Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, 2007.

SANTOS, D., SARMENTO, L. “O projecto AC/DC: acesso a corpora/disponibilização de corpora”. In Amália Mendes; Tiago Freitas (ed). Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002). Lisboa: APL, 2002.

SANTOS, D.; CARVALHO, P.; FREITAS, C.; OLIVEIRA, H. G. “Segundo HAREM: Directivas de anotação”. In: Mota, Cristina & Santos, Diana (orgs). Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: o Segundo HAREM. 2008.

SANTOS, D., MOTA, C. “Experiments in human-computer cooperation for the semantic annotation of portuguese corpora”. In: Calzolari et al (eds) Proceedings of LREC 2010. European Language Resources Association, 2010.

SANTOS, D.; MARQUES, R.; FREITAS, C.; SIMÕES, A.; MOTA, C. “Comparando anotações linguísticas na Gramateca: filosofia, ferramentas e exemplos”. In Domínios de Linguagem, 9(2), 11-26, 2015.

SANTOS, D.; BICK, E.; WLODEK, M. "Avaliando entidades mencionadas na coleção ELTeC-por". Linguamática 12.2, pp. 29-49, 2020.

SANTOS, D. ; ALVES, D. ; AMARO, R. ; BRANCO, I. A. ; FIALHO, O. ; FREITAS, C. ; HIGUCHI, S.; LANGFELDT, M. ; LOPES, J. M. ; PIRES, E. ; RAMOS, B. ; SANCHES, D. ; FUAO, R. S. ; PEREIRA, P. ; TERRA, P. “Leitura distante em português: resumo do Primeiro Encontro”. In MATLIT: MATERIALIDADES DA LITERATURA, v. 8, p. 279-298, 2020.

SAPIR, E. Language, an introduction to the study of speech. New York, Harcourt, 1949.

SARAWAGI, S. Information Extraction. Foundation and Trends in Databases, Vol. 1, No. 3 (2007) 261–377, DOI: 10.1561/1500000003.

SARDINHA, T. B. Lingüística de Corpus: histórico e problemática. DELTA, São Paulo, v. 16, n. 2, p. 323-367, 2000. https://doi.org/10.1590/S0102-44502000000200005. Acesso em 11/02/2020.

SARMENTO, L.; PINTO, A. S.; CABRAL, L. “REPENTINO – A wide-scope gazetteer for entity recognition in Portuguese”. In Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006 (PROPOR’2006). Springer Verlag, Berlim/Heidelberg, 13-17 de Maio de 2006, p. 31–40, 2006.

SAUSSURE, F. de. Curso de linguística geral. Organizado por Charles Bally e Albert Sechehaye. Prefácio de Isaac Nicolau Salum. 24. ed. São Paulo: Cultrix, 2002.

SCHNAPP, J., PRESNER, T., LUNENFELD, P., et al.: Digital humanities manifesto 2.0. Hentet 10, 2016 (2009).

SEEFELDT, D.; Thomas, W. G. “What is digital history?”. Intersections: History and New Media. Perspectives on History, 2009. Disponível em: https://www.historians.org/publications-and-directories/perspectives-on-history/may-2009/what-is-digital-history. Acesso em 25/10/2019.

SINCLAIR, J. “Corpus and text – basic principles”. In: WYNNE, M (ed.). Developing Linguistic Corpora: a Guide to Good Practice, 2005.

SMITH, A.; OSBORNE, M. “Using gazetteers in discriminative information extraction”. In Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X '06). Association for Computational Linguistics, Stroudsburg, PA, USA, 133-140, 2006.

SPARCK-JONES, K. Natural language processing: a historical review. in Artificial Intelligence Review. University of Cambridge, p. 2-10, 2001.

STRAKA, M.; STRAKOVA, J. UDPipe. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, 2016.

TABA, L. S. Extração automatica de relações semânticas do português. Dissertação. Universidade Federal de São Carlos, 2013.

TANAKA, S. "Pasts in a digital age." Writing History in the Digital Age, 2013. Disponível: https://www.jstor.org/stable/pdf/j.ctv65sx57.8.pdf. Acesso: 18/02/2021.

THESSEN, A.; CUI, H.; MOZZHERIN, D. Applications of Natural Language Processing in Biodiversity Science. Advances in bioinformatics. 2012.

VIEIRA, R.; GONÇALVES, P. N., SOUZA, J. G. C. de. "Processamento computacional de anáfora e correferência". Revista de Estudos da Linguagem 16.1, 2008: 263-284.

VILLAVICENCIO, A.; RAMISCH, C.; MACHADO, A.; CASELI, H. & FINATTO, M. J. Identificação de expressões multipalavra em domínios específicos. Linguamática, 2(1), 15-33, 2010.

WEIKUM, G.; THEOBALD, M. From information to knowledge: harvesting entities and relationships from web sources. In: Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. 2010. p. 65-76.

WIEDEMANN, G.; NIELKER, A. “Hands-on: A five day text mining course for humanists and social scientists in R”. Proceedings of the 1st Workshop on Teaching NLP for Digital Humanities, 2017, Berlin.

WILSON, A.; THOMAS, J. “Semantic Annotation”. In: GARSIDE et al. Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman. 1997

WITTGENSTEIN, Ludwig. Investigações Filosóficas. 2ª ed. Tradução de José Carlos Bruni. São Paulo, Abril Cutural (Os Pensadores), 1979.

WYNNE, M. Developing Linguistic Corpora: a Guide to Good Practice. 2005. Disponível online em http://ota.ox.ac.uk/documents/creating/dlc/index.htm. Acesso em 19/06/2017.

9
Anexos

Anexo 1

Questões enviadas por seis pesquisadores e/ou acadêmicos que utilizam o DHBB com frequência, face à consulta feita sobre que perguntas eles gostariam de ver respondidas automaticamente, caso fosse possível:

- Quais os políticos que nasceram antes da década de 1960, tiveram formação militar e ocuparam algum cargo no Executivo?
- Considerando a elite política de indivíduos nascidos entre 1900 e 1950 que ocuparam cargos no Executivo, como podemos observar a tendência e comportamento da variável "educação" ao longo do tempo?
- Como evolui a formação superior (ou média) dos quadros políticos durante a República Velha (1889-1930), 1ª Era Vargas (1930-1945), Democracia de 46 (1945-1964), Regime Militar (1964-1985) e Nova República (1985-)?
    - Há alguma diferença significativa entre ocupantes de cargos no Legislativo (deputados federais e senadores) e no Executivo (presidentes, governadores, ministros, prefeitos de capitais etc.)?
    - Alguma variação regional importante: Sul x Sudeste x Nordeste etc.?
- Quanto ao tema do recrutamento e da mobilidade da classe política – também agrupando pelos grandes períodos –, quais as trajetórias mais frequentes entre os ocupantes dos principais cargos no Executivo: presidentes e governadores? Quais os percentuais daqueles que passaram antes pelo Legislativo (locais e Federal) e/ou pelos Executivos locais? Muda muito quando passamos de um período a outro? De um estado a outro da Federação?
    - E outras diferenças sociais importantes, como gênero, ou cor da pele? Alguma alteração significativa de padrões?
- Quanto aos militares, como evolui sua presença na elite política? Como ingressaram na política: por via burocrática ou revolucionária?

- Qual o perfil educacional dos ministros do Superior Tribunal Militar, entre 1934 e 2010?
- Quantos ministros do STM trabalharam na Justiça Militar antes de serem nomeados para o cargo de ministro?
- Qual a idade dos ministros do Supremo Tribunal Federal ao serem nomeados?
- Quais os partidos políticos dos deputados federais e dos senadores, entre 1985 e 2010? Os parlamentares mudam muito de partidos, ao longo da sua trajetória, assim a resposta teria que contemplar esse aspecto também.
- Qual o perfil partidário dos ministros do Poder Executivo republicano?
- A qual(is) partido(s) cada político (digamos deputados federais) pertenceu? E quanto à carreira política, que cargos eletivos ocupou? Nos dois casos, a variável tempo seria fundamental.
- Quantos padres e pastores(as) exerceram - e exercem – cargos de deputados e senadores no Brasil da década de 1980 para cá? Incluir também sacerdote e reverendo.
- Quem são os políticos que detêm vínculos familiares com outros políticos? Que vínculos são esses?
- Sobre trajetória parlamentar desde 1946: informações sobre os parlamentares com seus respectivos cargos e partidos e se exerceram cargos executivos. O ideal é que o robô identificasse tipos de trajetória: se ascendente (vereador -> deputado estadual -> deputado federal -> senador), descendente ou não linear. Outra coisa interessante seria a produção de um gráfico de redes com as relações entre os verbetados, por exemplo, os mais citados apareceriam como nós maiores.

Anexo 2

Distribuição das formações acadêmicas presentes no DHBB:

Área de formação	Ocorrências
direito	3851
formação-militar	1082
engenharia	912
medicina	839
economia	633
administração	613
filosofia	289
contabilidade	192
letras	190
ciências-sociais	147
teologia	133
agronomia	121
educação	114
história	114
humanidades	100
ciências-políticas	86
química	78
pedagogia	77
jornalismo	76
farmácia	73
física	72
comunicação	71
odontologia	71
matemática	70
relações-internacionais	70
políticas-públicas	67
geografia	66
psicologia	56
criminologia	39
educação-física	39
arquitetura	36
veterinária	32
estatística	22
ciências-naturais	19
geologia	19
informática	18
serviço-social	18
metalurgia	14
urbanismo	14
eletrotécnica	13
turismo	12

biologia	11
radiologia	11
belas-artes	8
cinema	7
enfermagem	4
nutrição	4
biblioteconomia	1

Anexo 3

Expressões criadas para os temas de extração, com o quantitativo de ocorrências recuperadas.

NASCIMENTO		
01	Frases de exemplo	«Moroni Bing Torgan» nasceu em Porto Alegre, no dia 10 de junho de 1956.
		«Álvaro Francisco de Sousa» nasceu no dia 28 de fevereiro de 1903.
	Expressão	[classe="bio.*" & dicionario="dhbb" & pos="PROP.*"]+ [: pos!="PROP.*" :][]{0,1} [lema="nascer" & word!="nascido|nascer"] [pos="PRP.*"][]{0,21} [pos="NUM.*|ADJ.*"] [word="de"]? [pos="N.*"]? [word="de"]? [pos="NUM.*"]? [pos="PU"]
	Ocorrências	6.464

PERFIL EDUCACIONAL		
01	Exemplo(s)	iniciou o curso de economia na Universidade Federal de Pernambuco (Ufpe)
		concluiu o curso de engenharia eletrônica no Instituto de Tecnologia da Aeronáutica
	Sintaxe de busca	[dicionario="dhbb" & classe="biográfico" & lema="concluir|fazer|iniciar|ingressar|começar|completar|realizar"] []{0,10} [pos="DET.*"]? [lema="curso|mestrado|bacharelado|doutorado|pós-graduação"][]* "\."
	Ocorrências	3884
02	Exemplo(s)	frequentou a residência médica no Hospital Osvaldo Cruz, também na capital pernambucana.
		Frequentando também os cursos de filosofia pura e de administração da USP, graduou-se nos três cursos em 1972.
	Sintaxe de busca	[dicionario="dhbb" & classe="biográfico" & lema="frequentar"] [pos="DET.*|ADV" & word!="seu|sua|uma"] [pos="N|PROP" & word!="relação|sindicato|casa"][]* "\."
	Ocorrências	53
03	Exemplo(s)	formou-se em engenharia civil pela Universidade do Brasil
		Formou-se técnico em contabilidade no Colégio Duque de Caxias
	Sintaxe de busca	[dicionario="dhbb" & classe="biográfico" & lema="formar.*" & word!=".?ormadas|.?ormavam-se|.?ormara|.?ormaram|.?ormavam|.?ormaria|.?orma|.?ormei-me|.?ormassem|.?ormaram-se"] [word="bacharel|técnico"]? [pos!="DET.*" & word!="para|por|pel[oa]s|pelo|e|a|de|através|sobretudo|naquel.?|nesse|dessa|com|assim|.*mente|chapa|maioria.?|militantes"][pos="N" & word!="coligação|torno|aliança|união|gabinete|total|parlamentares|governo|maioria|fins|oposição|atividades|âmbito|Câmara|Congresso"] [pos="ADJ.*"]? []* "\."
	Ocorrências	1173
04	Exemplo(s)	bacharelando-se pela Faculdade de Direito da Universidade de Minas Gerais
		diplomou-se como engenheiro naval em 1964, na Escola Nacional de Engenharia

	Sintaxe de busca	[word="se"]? [dicionario="dhbb" & classe="biográfico" & lema="especializar.*|diplomar.*|bacharelar.*|graduar.*|doutorar.*" & word!= "especializad.?s|especializada"] []* "\."
	Ocorrências	2890
05	Exemplo(s)	ingressou na Faculdade de Direito da Universidade de Minas Gerais (UMG) .
		Ingressou em 1970 no curso de educação física da Comissão de Desportos da Aeronáutica no Rio de Janeiro (RJ) , formando-se em 1972.
	Sintaxe de busca	([dicionario="dhbb" & classe="biográfico" & lema="ingressar"])(([word="na"] [word="então|antiga|.?cademia|École|.?scola|Escuela|ESG|.?aculdade|Força|graduação|Johns|Marinha|Pontifícia|pós-graduação|School|Ufba|UFMA|.?niversidade"]|[word="no"][word="Ateneu|Colégio|.?urso|doutorado|ensino|Escola|Externato|First|.?inasial|ginásio|Instituto|internato|Liceu|mestrado|Pontifíci.?|Programa|.?eminário|Superior"]|[word="ano|mês"][word="seguinte"]? [pos="ADJ|PRP"][pos="PRP.*"][word="Faculdade|Escola|curso|mestrado|Colégio"])|([word="em"][pos="N|NUM.*"])([word="na|no"][pos="N|PROP" & sema="tit|occ|act.*|foreign_inst|top|inst.*|org.*" & word!="The|Movimento|Conselho|Tribunal|Ministério|Banco|Departamento|PSP|Empresa|Associação|Polícia|Ordem|Justiça|Companhia|Federação"]|[word="de"][pos="NUM.*|N"][word!="para|nos"][pos!="PU" & sema!="act-s_HH|sem_geom_act_Labs_inst|party|Lpath" & word!="Clube|.?erviço|Estado.*"])|("a" "seguir" "na|no" [pos="PROP|N" & word!="magistratura|carreira|Ação|Partido"]|"aos" []{0,2}"na|no" @[pos="PROP"]))[]* "\."
	Ocorrências	1786
06	Exemplo(s)	Estudou no Ginásio Diocesano de São Paulo e bacharelou-se em 1910 pela Faculdade de Ciências Jurídicas e Sociais.
		estudou medicina na Universidade Federal do Rio de Janeiro (UFRJ) de 1968 a 1973.
	Sintaxe de busca	[dicionario="dhbb" & classe="biográfico" & lema="estudar" & word!="estuda.?a|estudad.*"][pos!="DET.*"] @[pos="PERS.*|ADJ|NUM.*|PRP.*|PROP|KC|PU|N"][]* "\."
	Ocorrências	838
07	Exemplo(s)	cursou a Escola de Aperfeiçoamento de Oficiais.
		cursou a pós-graduação em engenharia nuclear e engenharia de produção.
	Sintaxe de busca	[dicionario="dhbb" & classe="biográfico" & lema="cursar"][pos=".*"][]* "\."
	Ocorrências	1506
08	Exemplo(s)	matriculou-se em Belo Horizonte na Faculdade Livre de Direito.
		matriculou-se na Universidade Federal da Bahia (UFBA) , graduando-se em engenharia elétrica em 1969.
	Sintaxe de busca	[word="se"]?[dicionario="dhbb" & classe="biográfico" & lema="matricular.*" & word!="matriculad.?s"][]* "\."
	Ocorrências	486
09	Exemplo(s)	mestre em ciências políticas e sociais, pela Escola de Ciências Políticas e Sociais da Universidade de Ottawa.

		Graduado em economia pela Faculdade de Ciências Econômicas da Universidade Federal do Ceará (UFC) , fez mestrado em administração financeira pela Universidade Federal do Rio de Janeiro (UFRJ) .
	Sintaxe de busca	[dicionario="dhbb" & classe="biográfico" & word=".?estre|.?outor|.?acharel|.?raduado|.?ós-graduado"] @[pos="PRP.*" & word!="de|dos|a|e|como|para|com"] [word!="honoris|posto|.?oja|mais|siderurgia|Exército|sarcasmo|prefeitura|brasilidade"][]* "\."
	Ocorrências	445
10	Exemplo(s)	obteve o título de doutor em ciências jurídicas através do concurso para catedrático da cadeira de economia política na Universidade Federal de Goiás.
		obteve o título de mestre em engenharia de produção pela COPPE / UFRJ e o de doutor em economia pelo University College of London (UCL) , na Inglaterra.
	Sintaxe de busca	[dicionario="dhbb" & classe="biográfico" & lema="obter"] [pos="DET.*"] @[word="doutorado|título|diploma"][]* "\."
	Ocorrências	96
11	Exemplo(s)	Sentou praça em abril de 1943,
		Sentou praça na Academia Militar das Agulhas Negras.
	Sintaxe de busca	[dicionario="dhbb" & classe="biográfico" & lema="sentar"] [word="praça"] []* [pos="PU"]
	Ocorrências	402

Relações familiares:

RELAÇÕES VÁLIDAS PARA O BIOGRAFADO: COM OUTRO BIOGRAFADO DO DHBB		
01	Exemplo	Paulo Maluf, seu padrinho
	Sintaxe	[dicionario="dhbb" & entidade="dhbb.*"]+ [: entidade!="dhbb.*" :] "," @[word="seu|sua"][sema="familia:lacos.*" & word!="companheiro.*|família.*"]
	Ocorrências	23
02	Exemplo	seu primo Olegário Maciel
	Sintaxe	[dicionario="dhbb" & pos="DET.*"][sema="familia:lacos.*" & word!="companheiro.*|família.*"] ","? [entidade="dhbb.*"]+ [: entidade!="dhbb.*" :]
	Ocorrências	655
03	Exemplo	filho de Antônio Carlos Ribeiro de Andrada
	Sintaxe	[dicionario="dhbb" & pos!="PROP.*"] "," [sema="familia:lacos.*" & word!="companheiro.*|família.*"] "de" [entidade="dhbb.*"]+ [: entidade!="dhbb.*" :]
	Ocorrências	304
04	Exemplo	filho do senador Benedito de Lira

	Sintaxe	[dicionario="dhbb" & pos!="PROP.*"] "," [sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="PRP.*"] [pos="N|ADJ"] [pos="ADJ"]? [entidade="dhbb.*"]+ [: entidade!="dhbb.*" :]
	Ocorrências	27
05	Exemplo	filha de Rubens Furlan, prefeito
	Sintaxe	[dicionario="dhbb" & sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="PRP"] [entidade="dhbb.*"]+ [: entidade!="dhbb.*" :] ".*" [pos="N"]
	Ocorrências	87
06	Exemplo	Seu irmão, o diplomata José de Paula Rodrigues Alves
	Sintaxe	[dicionario="dhbb" & pos!="PROP.*"][pos="DET.*"] [sema="familia:lacos.*" & word!="companheiro.*|família.*"] ","? [pos="DET.*"]? [pos="N|ADJ.*"] [pos="N|ADJ.*"]? [entidade="dhbb.*"]+ [: entidade!="dhbb.*" :]
	Ocorrências	60
07	Exemplo	filhos, entre os quais Adir Fiúza de Castro
	Sintaxe	[dicionario="dhbb" & sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="PU"] [pos="PRP"] [pos="SPEC_rel.*|PERS"] [pos=".*"] [pos="PU"]? [entidade="dhbb.*"]+ [: entidade!="dhbb.*" :]
	Ocorrências	18
08	Exemplo	Pedro Ludovico Teixeira, seu primo
	Sintaxe	[dicionario="dhbb" & entidade="dhbb.*"]+ [: entidade!="dhbb.*" :] "," [pos="DET.*"][sema="familia:lacos.*" & word!="companheiro.*|família.*"]
	Ocorrências	23

RELAÇÕES VÁLIDAS PARA O BIOGRAFADO: COM OUTRA PESSOA (COM NOME PRÓPRIO)		
01	Exemplo	filho de Antônio Carlos Ribeiro de Andrada e de Adelaide Feliciano Duarte de Andrada.
	Sintaxe	[dicionario="dhbb" & sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="PRP"] [pos="PROP.*"]+ [: pos!="PROP.*" :] "e" [pos="PRP"]? [pos="PROP.*"]+ [: pos!="PROP.*" :]
	Ocorrências	5.043
02	Exemplo	filho do cearense José Euclides Ferreira Gomes Filho
	Sintaxe	[dicionario="dhbb" & pos!="PROP.*"] "," [sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="PRP.*"] [pos="N|ADJ"] [pos="ADJ"]? [pos="PROP.*"]+ [: pos!="PROP.*" :]
	Ocorrências	496
03	Exemplo	Seu irmão, Júlio Campos
	Sintaxe	[dicionario="dhbb" & pos="DET.*"][sema="familia:lacos.*" & word!="companheiro.*|família.*"] ","? [pos="PROP.*"]+ [: pos!="PROP.*" :]
	Ocorrências	2.380

04	Exemplo	filho de Milton Leite da Silva, vereador por São Paulo
	Sintaxe	[dicionario="dhbb" & sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="PRP"] [pos="PROP.*"]+ [: pos!="PROP.*" :] ".*" [pos="N"]
	Ocorrências	445
05	Exemplo	ex-mulher de Jáder, Elcione Barbalho, concorreu à prefeitura de Belém.
	Sintaxe	[dicionario="dhbb" & sema="familia:lacos.*" & word!="companheiro.*|família.*"] "de" [pos="PROP.*"]+ [: pos!="PROP.*" :]","? [pos="PROP.*"]+ [: pos!="PROP.*" :]"," [lema="ser|concorrer"] [pos=".*" & lema!="criar|expulsar|acusar"] [pos="N"]?
	Ocorrências	7
06	Exemplo	Seu avô, o jurista Djalma Marinho
	Sintaxe	[dicionario="dhbb" & pos!="PROP.*"][pos="DET.*"] [sema="familia:lacos.*" & word!="companheiro.*|família.*"] ","? [pos="DET.*"]? [pos="N|ADJ.*"] [pos="N|ADJ.*"]? [pos="PROP.*"]+ [: pos!="PROP.*" :]
	Ocorrências	184
07	Exemplo	Seu pai, também conhecido como Ratinho
	Sintaxe	[dicionario="dhbb" & pos="DET.*"][sema="familia:lacos.*" & word!="companheiro.*|família.*"][pos="PU"] [pos=".*"] "conhecido" []* [pos="PROP.*"]+ [: pos!="PROP.*" :]
	Ocorrências	3
08	Exemplo	irmão do presidente, Pedro Collor
	Sintaxe	[dicionario="dhbb" & sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="PRP.*"] [pos="N"] [pos="ADJ"]? "," [pos="PROP.*"]+ [: pos!="PROP.*" :]
	Ocorrências	62
09	Exemplo	quatro filhos, entre os quais, Rodrigo de Castro
	Sintaxe	[dicionario="dhbb" & sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="PU"] [pos="PRP"] [pos="SPEC_rel.*|PERS"] [pos=".*"] [pos="PU"]? [pos="PROP.*"]+ [: pos!="PROP.*" :]
	Ocorrências	63
10	Exemplo	Daniela Santana Amorim, sua filha
	Sintaxe	[dicionario="dhbb" & sema="hum.*"]+ [: sema!="hum.*" :] "," [pos="DET.*"][sema="familia:lacos.*" & word!="companheiro.*|família.*"]
	Ocorrências	116
11	Exemplo	Ivo Gomes, outro de seus irmãos
	Sintaxe	[dicionario="dhbb" & pos="PROP.*"]+ [: pos!="PROP.*" :] "," [lema="DET.*"][pos="PRP.*"][pos="DET.*"][sema="familia:lacos.*" & word!="companheiro.*|família.*"]
	Ocorrências	5

RELAÇÕES VÁLIDAS PARA O BIOGRAFADO: COM OUTRA PESSOA (SEM NOME PRÓPRIO)		
01	Exemplo	Seu pai foi eleito deputado constituinte
	Sintaxe	[dicionario="dhbb" & pos="DET.*"][sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="V.*"] [lema="ser"]? [lema="eleger"] [pos="N"]
	Ocorrências	13
02	Exemplo	Seu pai foi prefeito e vereador
	Sintaxe	[dicionario="dhbb" & pos="DET.*"][sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="V.*"]? [pos="N" & sema!="familia:lacos.*"]
	Ocorrências	571
03	Exemplo	Seus dois filhos homens atuaram na política mineira
	Sintaxe	[dicionario="dhbb" & pos="DET.*"] [pos="N|NUM.*"]? [sema="familia:lacos.*" & word!="companheiro.*|família.*"][pos="N"]? [lema="atuar"][]{0,5}[lema="política"]
	Ocorrências	4
04	Exemplo	Seu pai, também diplomata
	Sintaxe	[dicionario="dhbb" & pos="DET.*"] [sema="familia:lacos.*" & word!="companheiro.*|família.*"]"," [pos="ADV.*"] [pos="N|ADJ"] [pos="ADJ"]?
	Ocorrências	37

RELAÇÕES VÁLIDAS ENTRE TERCEIROS (ENTRE BIOGRAFADOS DO DHBB)		
01	Exemplo	Francisco Dornelles, sobrinho de Tancredo Neves
	Sintaxe	[dicionario="dhbb" & entidade="dhbb.*"]+ [: entidade!="dhbb.*" :] "," [sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="PRP.*"] [pos="N"]?[]{0,3}[entidade="dhbb.*"]+ [: entidade!="dhbb.*" :]
	Ocorrências	31
02	Exemplo	Amaral Peixoto e seu genro, Wellington Moreira Franco
	Sintaxe	[dicionario="dhbb" & entidade="dhbb.*"]+ [: entidade!="dhbb.*" :] "e" [pos="DET.*"] [sema="familia:lacos.*|família.*" & word!="companheiro.*"] [pos="PRP.*"]? "," [entidade="dhbb.*"]+ [: entidade!="dhbb.*" :]
		2
03	Exemplo	primo de Jorge Vargas, José Israel Vargas
	Sintaxe	[dicionario="dhbb" & sema="familia:lacos.*" & word!="companheiro.*|família.*"] "de" [entidade="dhbb.*"]+ [: entidade!="dhbb.*" :] "," [entidade="dhbb.*"]+ [: entidade!="dhbb.*" :]
	Ocorrências	2

RELAÇÕES VÁLIDAS ENTRE TERCEIROS (ENTRE PESSOAS COM NOME PRÓPRIO)		
01	Exemplo	Luís Fernando Zoghbi, filho de João Carlos Zoghbi

	Sintaxe	[dicionario="dhbb" & sema="hum.*"]+ [: sema!="hum.*" :]"," [sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="PRP.*"] [pos="N"]?[]{0,3}[pos="PROP.*"]+ [: pos!="PROP.*" :]
	Ocorrências	380
02	Exemplo	Talvane Albuquerque e sua irmã, Teresinha Albuquerque
	Sintaxe	[dicionario="dhbb" & pos="PROP.*"]+ [: pos!="PROP.*" :] "e" [pos="DET.*"] [sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="PRP.*"]? "," [pos="PROP.*"]+ [: pos!="PROP.*" :]
	Ocorrências	13
03	Exemplo	esposa de Davi, Maria Augusta de Oliveira
	Sintaxe	[dicionario="dhbb" & sema="familia:lacos.*" & word!="companheiro.*|família.*"] "de" [pos="PROP.*"]+ [: pos!="PROP.*" :] "," [pos="PROP.*"]+ [: pos!="PROP.*" :]
	Ocorrências	56
04	Exemplo	Sua esposa era prima de José Aparecido de Oliveira
	Sintaxe	[dicionario="dhbb" & pos="DET.*"] [sema="familia:lacos.*" & word!="companheiro.*|família.*"] [lema="ser"] [sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="PRP"] [pos="PROP.*"]+ [: pos!="PROP.*" :]
	Ocorrências	3
05	Exemplo	deputado estadual Marco Antônio Alencar, filho de Marcelo
	Sintaxe	[dicionario="dhbb" & pos="N"] [pos="ADJ"]? [pos="PROP.*"]+ [: pos!="PROP.*" :] []{0,5} [pos="PU"]+ [sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="PRP"] [pos="PROP.*"]+ [: pos!="PROP.*" :]
	Ocorrências	72
06	Exemplo	Leda Collor, mãe do presidente
	Sintaxe	[dicionario="dhbb" & pos="PROP.*"]+ [: pos!="PROP.*" :] "," [sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="PRP.*"] [pos="N"] [pos="ADJ"]?
	Ocorrências	300
07	Exemplo	irmão mais velho de Collor, Leopoldo
	Sintaxe	[dicionario="dhbb" & sema="familia:lacos.*" & word!="companheiro.*|família.*"] [pos="ADV.*"] [pos="ADJ"] [pos="PRP"][pos="PROP.*"]+ [: pos!="PROP.*" :] "," [pos="PROP.*"]+ [: pos!="PROP.*" :]
	Ocorrências	2

Anexo 4

Cargos que justificam a inclusão de determinado personagem no DHBB.

Poder Legislativo

dep. fed.	Deputado Federal
dep. fed. prof.	Deputado Federal Profissional
sen.	Senador
const.	Constituinte

Poder Executivo

pres. Rep.	Presidente da República
pres. Rep. eleito	Presidente da República Eleito (foram eleitos, mas não tomaram posse)
vice-pres. Rep.	Vice-presidente da República
sec. Pres. Rep.	Secretário da Presidência da República
ch. Gab. Civ. Pres. Rep.	Chefe de Gabinete Civil da Presidência da República
ass. econ. Pres. Rep.	Assessoria Econômica da Presidência da República
ass. Pres. Rep.	Assessor da Presidência da República
aux. gab. Pres. Rep.	Auxiliar de Gabinete da Presidência da República
consult. ger. Rep.	Consultor geral da República
prim.-min.	Primeiro-ministro
pres. junta Gov. Prov.	Presidente da junta governativa provisória
junta gov.	Junta governativa
junta gov. prov.	Junta governativa provisória
junta mil.	Junta governativa militar
ch. gov. prov.	Chefe do governo provisório
interv.	Interventor federal
gov.	Governador
gov. mil.	Governador militar
gov. prov.	Governador Provisório
vice-gov.	Vice-governador
pref.	Prefeito
min. Ação Social	Ministro da Ação Social
min. Adm.	Ministro da Administração
min. Adm. Ref. Est.	Ministro da Administração Federal e Reforma do Estado
min. Aer.	Ministro da Aeronáutica
min. Agric.	Ministro da Agricultura
min. Agric. Abast. e Ref. Agr.	Ministro da Agricultura, do Abastecimento e da Reforma Agrária
min. Amazônia Legal e Meio Ambiente	Ministro da Amazônia Legal e do Meio Ambiente
min. Assist. e Prom. Social	Ministro da Secretaria Especial da Assistência e Promoção Social
min. Assist. Social	Secretária de Assistência Social (com status de ministro)
min. Assunt. Estrat.	Ministro da Secretaria de Assuntos Estratégicos (SAE)
min. Bem-Estar Soc.	Ministro do Bem-Estar Social
min. Casa Civ. Pres. Rep.	Ministro da Casa Civil da Presidência da República
min. Casa Mil. Pres. Rep.	Ministro da Casa Militar da Presidência da República
min. CGU	Ministro da Controladoria-Geral da União
min. ch. EMFA	Ministro-chefe do Estado-Maior das Forças Armadas (EMFA)
min. ch. SAF	Ministro-chefe da Secretaria de Administração Federal (SAF)
min. ch. Secretaria de Governo	Ministro-chefe da Secretaria de Governo
min. ch. Seplan	Ministro-chefe da Secretaria de Planejamento (Seplan)
min. ch. SNI	Ministro-chefe do Serviço Nacional de Informações (SNI)
min. Cidades	Ministro das Cidades
min. Ciên. Tecn.	Ministro da Ciência e Tecnologia
min. Ciên. Tecn. e Inov.	Ministro da Ciência, Tecnologia e Inovação (MCTI)
min. Comunic.	Ministro das Comunicações
min. ch. Comunic. Pres. Rep.	Ministro-chefe da Secretaria de Comunicação da Presidência da República

min. Controle e Transparência	Ministro de Estado do Controle e da Transparência
min. Cultura	Ministro da Cultura
min. Defesa	Ministro da Defesa
min. Desenv. Agr.	Ministro do Desenvolvimento Agrário
min. Desenv. Ind. Com.	Ministro do Desenvolvimento, Indústria e Comércio
min. Desenv. Regional	Ministro do Desenvolvimento Regional
min. Desenv. Soc.	Ministro do Desenvolvimento Social
min. Desenv. Urb. e Meio Ambiente	Ministro do Desenvolvimento Urbano e Meio Ambiente (MDU)
min. Direitos Humanos	Ministro da Secretaria Especial de Direitos Humanos (SEDH)
min. Econ.	Ministro da Economia
min. Educ.	Ministro da Educação
min. Educ. e Cultura	Ministro da Educação e Cultura (MEC)
min. Educ. e Desp.	Ministro da Educação e do Desporto (MEC)
min. Esporte	Ministro do Esporte
min. Esporte e Turis.	Ministro do Esporte e Turismo
min. Exérc.	Ministro do Exército
min. Extr. Ass. Fundiários	Ministro Extraordinário para Assuntos Fundiários
min. Extr. Ass. Irrigação	Ministro Extraordinário para Assuntos de Irrigação
min. Extr. Coord. Pol.	Ministro Extraordinário para a Coordenação Política
min. Extr. Coord. Órgãos Reg.	Ministro Extraordinário para a Coordenação dos Organismos Regionais
min. Extr. Criança	Ministro Extraordinário da Criança
min. Extr. Desburoc.	Ministro Extraordinário da Desburocratização
min. Extr. Integr. América Latina	Ministro Extraordinário para Assuntos de Integração da América Latina
min. Extr. Pol. Fund. e Desenv. Agr.	Ministro Extraordinário de Política Fundiária e Desenvolvimento Agrário
min. Extr. Ref. Admin.	Ministro Extraordinário para Assuntos da Reforma Administrativa
min. Extr. Ref. Inst.	Ministro Extraordinário das Reformas Institucionais
min. Faz.	Ministro da Fazenda
min. Gab. Seg. Inst.	Ministro-chefe Gabinete de Segurança Institucional da Presidência da República (GSI)
min. Guerra	Ministro da Guerra
min. Hab. Urb. e Meio Amb.	Ministro de Habitação, Urbanismo e Meio-Ambiente (MHU)
min. Ind. Com. e Tur.	Ministro da Indústria, Comércio e Turismo
min. Ind. e Com.	Ministro da Indústria e do Comércio
min. Infraestrutura	Ministro da Infraestrutura
min. Integr. Nac.	Ministro da Integração Nacional
min. Integr. Reg.	Ministro da Integração Regional
min. Interior	Ministro do Interior
min. Interior e Just.	Ministro do Interior e Justiça
min. Just.	Ministro da Justiça
min. Mar.	Ministro da Marinha
min. Meio Amb. e Amazônia Legal	Ministro do Meio Ambiente e da Amazônia Legal
min. Meio Ambiente	Ministro do Meio Ambiente
min. Meio Ambiente, Recursos Hídricos e da Amazônia Legal	Ministro do Meio Ambiente, dos Recursos Hídricos e da Amazônia Legal
min. Minas e En.	Ministro das Minas e Energia
min. Orç.	Ministro de Orçamento e Gestão
min. Pesca	Ministro da Pesca
min. Planej.	Ministro do Planejamento
min. Planejamento, Orçamento e Gestão	Ministro do Planejamento, Orçamento e Gestão
min. Pol. Fund. e Desenv. Agr.	Ministro de Política Fundiária e Desenvolvimento Agrário
min. Prev. e Assist. Social	Ministro da Previdência e Assistência Social
min. Prev. Social	Ministro da Previdência e Assistência Social
min. Ref. e Desenv. Agr.	Ministro da Reforma e Desenvolvimento Agrário
min. Rel. Ext.	Ministro das Relações Exteriores
min. Rel. Inst.	Ministro das Relações Institucionais
min. Saúde	Ministro da Saúde

min. Sec. Ass. Estrat.	Ministro da Secretaria de Assuntos Estratégicos
min. Sec. Des. Urbano	Ministro da Secretaria Especial de Desenvolvimento Urbano
min. Sec. Esp. Pol. Mulheres	Ministro da Secretaria Especial de Política para as Mulheres (SPM)
min. Sec. Esp. Pol. Prom. Iguald. Racial	Ministro da Secretaria Especial de Políticas de Promoção da Igualdade Racial (SEPPIR)
min. Sec. Esp. Portos	Ministro da Secretaria Especial de Portos da Presidência da República (SEP/PR)
min. Sec. Especial de Aquicultura e Pesca	Ministro da Secretaria Especial de Aquicultura e Pesca
min. Sec. Ger. Pres. Rep.	Ministro da Secretaria Geral da Presidência da República
min. Sec. Micro e Pequena Empresa	Ministro da Secretaria da Micro e Pequena Empresa da Presidência da República
min. Sec. Pol. Region.	Ministro da Secretaria Especial de Políticas Regionais da Presidência da República
min. TCU	Ministro do Tribunal de Contas da União
min. Trab.	Ministro do Trabalho
min. Trab. Admin.	Ministro do Trabalho e Administração
min. Trab. e Emprego	Ministro do Trabalho e do Emprego
min. Trab. e Prev. Soc.	Ministro do Trabalho e da Previdência Social
min. Transp.	Ministro dos Transportes
min. Transp. e Com.	Ministro dos Transportes e Comunicações
min. Tur.	Ministro do Turismo
min. Tur. e Esp.	Ministro do Esporte e Turismo
min. Viação	Ministro da Viação e Obras Públicas
min. Viação e Agric.	Ministro da Viação e Obras Públicas

Militares

militar	Militar
adido mil.	Adido militar
insp. 1º Gr. RM	Inspetor do 1º Grupo de Regiões Militares
insp. 2º Gr. RM	Inspetor do 2º Grupo de Regiões Militares
comdo. naval Brasília	Comando Naval de Brasília
comte. 12ª RM	Comandante da 12ª Região Militar
comte. 1ª RM	Comandante da 1ª Região Militar
comte. 2ª RM	Comandante da 2ª Região Militar
comte. 3ª RM	Comandante da 3ª Região Militar
comte. 4ª RM	Comandante da 4ª Região Militar
comte. 5ª RM	Comandante da 5ª Região Militar
comte. 7ª RM	Comandante da 7ª Região Militar
comte. Art. FEB	Comandante da Artilharia Divisionária da FEB
comte. CFN	Comandante do Corpo de Fuzileiros Navais
comte. Comdo. Mil. Amazônia	Comandante do Comando Militar da Amazônia
comte. Comdo. Mil. Brasília	Comandante do Comando Militar de Brasília
comte. Comdo. Mil. Planalto	Comandante do Comando Militar do Planalto
comte. Depto. Eng. Comunic. Ex.	Comandante do Departamento de Engenharia e Comunicações do Exército
comte. EMA	Comandante do Estado-Maior da Armada (EMA)
comte. ESG	Comandante da Escola Superior de Guerra (ESG)
comte. FEB	Comandante da Força Expedicionária Brasileira (FEB)
comte. Força Naval NE	Comandante da Força Naval do Nordeste
comte. ger. PMDF	Comandante-geral da Polícia Militar do Distrito Federal
comte. ger. PMRJ	Comandante-geral da Polícia Militar do Rio de Janeiro
comte. I DN	Comandante do I Distrito Naval
comte. I Ex.	Comandante do I Exército
comte. I ZA	Comandante da I Zona Aérea
comte. II Comar	Comandante do II Comando Aéreo Regional
comte. II DN	Comandante do II Distrito Naval
comte. II Ex.	Comandante do II Exército
comte. II ZA	Comandante da II Zona Aérea
comte. III Comar	Comandante do III Comando Aéreo Regional
comte. III DN	Comandante do III Distrito Naval

comte. III Ex.	Comandante do III Exército
comte. III ZA	Comandante da III Zona Aérea
comte. Inf. FEB	Comandante de Infantaria da Força Expedicionária Brasileira
comte. IV Comar	Comandante do IV Comando Aéreo Regional
comte. IV DN	Comandante do IV Distrito Naval
comte. IV Ex.	Comandante do IV Exército
comte. IV ZA	Comandante da IV Zona Aérea
comte. V DN	Comandante do V Distrito Naval
comte. V ZA	Comandante da V Zona Aérea
comte. VI Comar	Comandante do VI Comando Aéreo Regional
comte. VI DN	Comandante do VI Distrito Naval
comte. VI ZA	Comandante da VI Zona Aérea
comte. Zona Mil. Centro	Comandante da Zona Militar Centro
comte. Zona Mil. Leste	Comandante da Zona Militar Leste
comte. Zona Mil. Norte	Comandante Zona Militar Norte
comte. Zona Mil. Sul	Comandante Zona Militar Sul
comte.-em-ch. Esquadra	Comandante-em-chefe da Esquadra
ch. VI Comar	Chefe do VI Comando Aéreo Regional (VI Comar)
ch. EM FEB	Chefe do Estado-Maior da Força Expedicionária Brasileira
ch. EM Geral	Chefe do Estado-Maior Geral
ch. EM Gov. Prov.	Chefe do Estado-Maior do Governo Provisório
ch. EM Pres. Rep.	Chefe do Estado-Maior da Presidência da República
ch. Depto. Eng. Comunic. Ex.	Chefe do Departamento de Engenharia e Comunicações do Exército
ch. Depto. Ens. Pesq. Ex.	Chefe do Departamento de Ensino e Pesquisa do Exército
ch. Depto. Ger. Admin. Ex.	Chefe do Departamento Geral de Administração do Exército
ch. Depto. Ger. Pess. Ex.	Chefe do Departamento Geral de Pessoal do Exército
ch. Depto. Ger. Serv. Ex.	Chefe do Departamento Geral de Serviços do Exército
ch. Depto. Mat. Bél. Ex.	Chefe do Departamento de Material Bélico do Exército
ch. Depto. Prod. Obras Ex.	Chefe do Departamento de Produção de Obras do Exército
ch. Depto. Prov. Ger. Ex.	Chefe do Departamento de Provisão Geral do Exército
ch. Depto. Téc. Prod. Ex.	Chefe do Departamento Técnico de Produção do Exército
ch. EMA	Chefe do Estado-Maior da Armada
ch. Emaer	Chefe do Estado-Maior da Aeronáutica
ch. EME	Chefe do Estado-Maior do Exército
ch. EMFA	Chefe do Estado-Maior das Forças Armadas
ch. Gab. Mil. Pres. Rep.	Chefe do Gabinete Militar da Presidência da República

Burocracia Pública

pres. Cons. Nac. Café	Presidente do Conselho Nacional do Café
pres. TCU	Presidente do Tribunal de Contas da União
pres. Telebrás	Presidente da Telebrás
pres. SNA	Presidente da Sociedade Nacional de Agricultura
pres. Petrobras	Presidente da Petrobrás
pres. IAPC	Presidente do Instituto de Aposentadoria e Pensões dos Comerciários
pres. IBC	Presidente do Instituto Brasileiro do Café
pres. INPS	Presidente do Instituto Nacional de Previdência Social
pres. IPASE	Presidente do Instituto de Previdência e Assistência aos Servidores do Estado
pres. Light	Presidente da Light
pres. CVRD	Presidente da Companhia Vale do Rio Doce
pres. DASP	Presidente do Departamento Administrativo do Serviço Público
pres. DNC	Presidente do Departamento Nacional do Café
pres. Eletrobrás	Presidente das Centrais Elétricas Brasileiras S.A.
pres. CNP	Presidente do Conselho Nacional de Petróleo
pres. CNPq	Presidente do Conselho Nacional de Pesquisas
pres. BNH	Presidente do Banco Nacional da Habitação
pres. Bco. Bras.	Presidente do Banco do Brasil
pres. Bco. Central	Presidente do Banco Central
pres. BNDE	Presidente do Banco Nacional de Desenvolvimento Econômico
pres. BNDES	Presidente do Banco Nacional do Desenvolvimento Econômico e Social
dir. DASP	Diretor do Departamento Administrativo do Serviço Público

dir. Depto. Nac. Prop.	Diretor do Departamento Nacional de Propaganda (DNP)
dir. Depto. Prop. Dif. Cult.	Diretor do Departamento de Propaganda e Difusão Cultural (DPDC)
dir. DIP	Diretor do Departamento de Imprensa e Propaganda
dir. ger. ABIN	Diretor-geral da Agência Brasileira de Inteligência
dir. Ag. Nac.	Diretor da Agência Nacional
dir. Av. Mil.	Diretor da Aviação Militar
ch. AERP	Chefe da Assessoria Especial de Relações Públicas
ch. oper. Petrobras	chefe de operações da Petrobras
superint. Sudene	Superintendente da Superintendência do Desenvolvimento do Nordeste
superint. Sumoc	Superintendente da Superintendência da Moeda e do Crédito
sec. Adm. Fed.	Secretário de Administração Federal
sec. Ass. Estratég.	Secretário de Assuntos Estratégicos (SAE)
sec. Ciên. Tecn.	Secretário de Ciência e Tecnologia
sec. Cons. Téc. Econ. Fin.	Secretário do Conselho Técnico de Economia e Finanças (CTEF)
sec. Coord. Pol.	Secretário da Secretaria de Coordenação Política
sec. Cultura	Secretário de Cultura
sec. esp. Agricultura e Pesca	Secretário da Secretaria Especial de Aquicultura e Pesca
sec. esp. Cons. Desenvol. Econ. Soc.	Secretário da Secretaria Especial do Conselho de Desenvolvimento Econômico e Social
sec. esp. Pol. Prom. Igualdade Racial	Secretário da Secretaria Especial de Políticas de Promoção da Igualdade Racial (SEPPIR)
sec. ger. ANL	Secretário-geral da Aliança Nacional Libertadora
sec. Impr. e Divulg. Pres. Rep.	Secretário na Secretaria de Imprensa e Divulgação da Presidência da República
sec. Nac. Comun.	Secretário da Secretaria Nacional de Comunicações
sec. Nac. Desportos	Secretário da Secretaria Nacional de Desportos
sec. Nac. Dir. Hum.	Secretário da Secretaria Nacional de Direitos Humanos
sec. Nac. Meio Ambiente	Secretário da Secretaria Nacional do Meio Ambiente
sec. Nac. Pol. Mulheres	Secretário da Secretaria Especial de Políticas para as Mulheres (SPM)

Segurança Pública

deleg. mil. SP	Delegado Militar de São Paulo
deleg. pol. SP	Delegado da Polícia de SP
dir. ger. DPF	Diretor-geral do Departamento de Polícia Federal
dir. ger. DFSP	Diretor-geral do Departamento Federal de Segurança Pública
ch. CGI	Chefe da Comissão Geral de Investigações
ch. CODI	Chefe do Centro de Operações de Defesa Interna
ch. DFSP	Chefe do Departamento Federal de Segurança Pública
ch. DPF	Chefe do Departamento de Polícia Federal
ch. pol. DF	Chefe de Polícia do Distrito Federal
ch. SNI	Chefe do Serviço Nacional de Informação

Trabalho e Indústra

pres. CNA	Presidente da Confederação Nacional da Agricultura
pres. ACRJ	Presidente da Associação Comercial do Rio de Janeiro
pres. CAT	Presidente da Central Autônoma dos Trabalhadores
pres. CGT	Presidente do Comando Geral dos Trabalhadores
pres. CGTB	Presidente da Central Geral dos Trabalhadores do Brasil
pres. CIB	Presidente do Centro Industrial do Brasil
pres. CIESP	Presidente do Centro das Indústrias do Estado de São Paulo
pres. CIRJ	Presidente do Centro Industrial do Rio de Janeiro
pres. CIRJ/FIDF	Presidente do Centro Industrial do Rio de Janeiro (CIRJ) / Presidente da Federação das Indústrias do Distrito Federal (FIDF)
pres. CIRJ/Fiega	Presidente do Centro Industrial do Rio de Janeiro (CIRJ) / Presidente da Federação das Indústrias do Estado da Guanabara (Fiega)
pres. CIRJ/FIRJ	Presidente do Centro Industrial do Rio de Janeiro (CIRJ) / Presidente da Federação das Indústrias do Rio de Janeiro (FIRJ)
pres. CNC	Presidente da Confederação Nacional do Comércio
pres. CNI	Presidente da Confederação Nacional da Indústria
pres. CNTC	Presidente da Confederação Nacional dos Trabalhadores no Comércio

pres. CNTI	Presidente da Confederação Nacional dos Trabalhadores na Indústria
pres. CNTTMFA	Presidente da Confederação Nacional dos Trabalhadores em Transportes Marítimos, Fluviais e Aéreos
pres. CNTTT	Presidente da Confederação Nacional dos Trabalhadores em Transportes Terrestres
pres. Contag	Presidente da Confederação Nacional dos Trabalhadores na Agricultura
pres. Contec	Presidente da Confederação Nacional dos Trabalhadores em Estabelecimentos de Crédito
pres. CUT	Presidente da Central Única dos Trabalhadores
pres. FIEGA/CIRJ	Presidente da Federação das Indústrias do Estado do Rio de Janeiro (Firjan) / Presidente do Centro Industrial do Rio de Janeiro (CIRJ)
pres. FIESP	Presidente da Federação das Indústrias do Estado de São Paulo
pres. FIESP/CIESP	Presidente da Federação das Indústrias do Estado de São Paulo (FIESP) / Presidente do Centro das Indústrias do Estado de São Paulo (CIESP)
pres. FIRJ	Presidente da Federação das Indústrias do Rio de Janeiro
pres. Firjan	Presidente da Federação das Indústrias do Estado do Rio de Janeiro
pres. Firjan-CIRJ	Presidente da Federação das Indústrias do Estado do Rio de Janeiro (Firjan) / Presidente do Centro Industrial do Rio de Janeiro (CIRJ)
pres. Força Sindical	Presidente da Força Sindical
pres. NCST	Presidente da Nova Central Sindical dos Trabalhadores
pres. SDS	Presidente da Social Democracia Sindical
pres. UGT	Presidente da União Geral dos Trabalhadores
dir. CGT	Diretor do Comando Geral dos Trabalhadores
líder seringueiro	líder seringueiro
sind.	Sindicalista

Partidos e movimentos políticos

pres. PTB	Presidente do Partido Trabalhista Brasileiro
pres. ANL	Presidente da Aliança Nacional Libertadora
atent. Toneleros	Atentado da rua Toneleros
col. Prestes	Coluna Prestes
contestado	Guerra do Contestado
líder guerrilheiro	líder guerrilheiro
líder soc. civil	Líder da sociedade civil
mov. comunista	Movimento comunista
mov. integralista	Movimento integralista
mov. tenentista	Movimento tenentista
rev.	Revolucionário
rev. 1892	Revolucionário de 1892
rev. 1922	Revolucionário de 1922
rev. 1923	Revolucionário de 1923
rev. 1924	Revolucionário de 1924
rev. 1925	Revolucionário de 1925
rev. 1926	Revolucionário de 1926
rev. 1930	Revolucionário de 1930
rev. 1932	Revolucionário de 1932
rev. 1935	Revolucionário de 1935
rev. 1938	Revolucionário de 1938
rev. 1964	Participante do Golpe de 1964
rev. Aragarças	Revolta de Aragarças
rev. Jacareacanga	Revolta de Jacareacanga
rev. marinheiros	Revolta dos Marinheiros
rev. Princesa	Revolta de Princesa
rev. sargentos	Revolta dos Sargentos
membro ANL	Membro da Aliança Nacional Libertadora

Magistratura e Judiciário

pres. TSE	Presidente do Tribunal Superior Eleitoral
min. TSE	Ministro do Tribunal Superior Eleitoral
min. STF	Ministro do Supremo Tribunal Federal

min. STJ	Ministro do Superior Tribunal de Justiça
min. STM	Ministro do Supremo Tribunal Militar; Ministro do Superior Tribunal Militar
pres. STF	Presidente do Supremo Tribunal Federal
adv. geral União	Advogado-geral da União
juiz TSN	Juiz do Tribunal de Segurança Nacional
jurista	Jurista
magistrada	Magistrada
magistrado	Magistrado
proc. ger. Rep.	Procurador-geral da República
promotor TSN	Procurador do Tribunal de Segurança Nacional

Esfera educacional

reitor UB	Reitor da Universidade do Brasil
reitor UDF	Reitor da Universidade do Distrito Federal
reitor UFRJ	Reitor da Universidade Federal do Rio de Janeiro
reitor UnB	Reitor da Universidade de Brasília
reitor Univ. RJ	Reitor da Universidade do Rio de Janeiro
reitor USP	Reitor da Universidade de São Paulo
pres. UNE	Presidente da União Nacional dos Estudantes
líder estudantil	líder estudantil

Esfera Religiosa

arceb.	Arcebispo
religioso	Religioso
sec. ger. CNBB	Secretário-geral da Conferência Nacional dos Bispos do Brasil
pres. CNBB	Presidente da Conferência Nacional dos Bispos do Brasil

Profissionais liberais

arquiteto	arquiteto
ecologista	ecologista
empresário	empresário
escritor	escritor
intelectual	intelectual
jornalista	jornalista
pensador político	pensador político
urbanista	urbanista

Outros

cand. pres. Rep.	Candidato à presidência da República
cand. vice-pres. Rep.	Candidato à vice-presidência da República
pres. Acad. Bras. Letras	Presidente da Academia Brasileira de Letras (ABL)
diplomata	diplomata
emb.	Embaixador
encar. neg.	Encarregado de negócios
ch. missão econ.	Chefe da missão econômica
pres. FGV	Presidente da Fundação Getúlio Vargas

Anexo 5

Quadro com as principais famílias que comandam ou comandaram a política em determinados estados e regiões do país, segundo matéria da edição impressa da Folha de S. Paulo de 19 de agosto de 2018[52].

Família e estado	Principal político	Principais familiares na política
Sarney (MA)	José Sarney • Desbancou Vitorino Freire, até então principal coronel político do Maranhão. Presidente da República de 1985 a 1990. Governou o Maranhão e foi senador pelo Maranhão e por Amapá. Atualmente está sem mandato. Tem 88 anos	Roseana Sarney | Filha • Governou o Maranhão e foi senadora. É candidata ao governo do estado Sarney Filho | Filho • Deputado federal com nove mandatos, foi ministro dos governos FHC e Temer. É candidato ao Senado Adriano Sarney | Primo • Deputado estadual, tenta reeleição
Calheiros (AL)	Renan Calheiros • Presidiu o Senado, foi ministro da Justiça no governo FHC. É candidato à reeleição no Senado	Olavo Neto | Sobrinho • Prefeito de Murici (AL) Renan Filho | Filho • Governador de Alagoas, candidato à reeleição Olavo Calheiros | Irmão • Deputado estadual, tenta reeleição Renildo Calheiros | Irmão • Filiado ao PC do B, tem atuação política em Pernambuco. É candidato a deputado federal
Collor (AL)	Fernando Collor • Filho do ex-senador e ex-governador Arnon de Melo, foi presidente da República de 1990 a 1992, ano em que sofreu impeachment. É candidato ao governo de Alagoas	Arnon de Melo (1911-1983) | Pai Durante uma briga no Senado com o adversário político Silvestre Péricles de Góis Monteiro, sacou o revólver em 1963, deu três tiros e acabou matando um outro senador, que nada tinha a ver com a história Fernando James | Filho Candidato a deputado federal
Barbalho (PA)	Jader Barbalho • Governou o Pará, foi deputado, ministro e é senador. É candidato à reeleição	Helder Barbalho | Filho Ex-prefeito, ex-ministro, é candidato ao governo do estado Elcione Barbalho | Ex-mulher • Deputada federal, tenta reeleição Simone Morgado | Ex-mulher • Deputada federal, tenta reeleição

[52] https://www1.folha.uol.com.br/poder/2018/08/dinastias-politicas-do-brasil-lancam-mais-de-60-candidatos-nas-eleicoes.shtml?cmpid=assmob&origin=folha . Acesso em 23/08/2019

Família e estado	Principal político	Principais familiares na política
		Priante | Primo • Deputado federal, tenta reeleição
Andrada (MG)	José Bonifácio de Andrada e Silva (1763-1838) • Conhecido como Patriarca da Independência, foi um dos principais políticos desse período	Bonifácio de Andrada | Descendente • Na Câmara desde 1979, relatou a segunda denúncia contra o presidente Michel Temer, com voto favorável ao emedebista. Não tentará a reeleição Lafayette Andrada | Filho de Bonifácio • Deputado estadual, é candidato a deputado federal Doorgal Andrada | Filho de Bonifácio • Vereador, é candidato a deputado estadual Toninho Andrada | Filho de Bonifácio • Candidato a deputado estadual
Neves/Cunha (MG)	Tancredo Neves (1910-1985) • Ministro de Getulio Vargas, primeiro ministro no pré-ditadura e primeiro civil eleito no pós-ditadura, pelo Congresso, morreu antes de tomar posse	Aécio Neves | Neto • Neto de políticos tanto por parte de mãe quanto por parte de pai, foi presidente da Câmara, governou Minas, tentou se eleger presidente da República e hoje é Senador. Um dos pivôs do escândalo da JBS, vai tentar, agora, se eleger deputado federal Andrea Neves | Neta • Foi desde sempre a principal auxiliar de Aécio
Arraes (PE)	Miguel Arraes (1916-2005) • Governador de Pernambuco, foi preso e deposto pela ditadura militar	Eduardo Campos (1965-2014) | Neto • Governador de Pernambuco, morreu em acidente aéreo quando disputava a Presidência, em 2014 Ana Arraes | Filha • Ministra do TCU Paulo Câmara | Casado com um parente • Casada com uma prima de Eduardo Campos, foi eleito governador de Pernambuco em nome do clã. Tenta a reeleição Marília Arraes | Neta • Vereadora no Recife, teve a candidatura ao governo barrada pelo PT. É candidata a deputada federal João Campos | Bisneto • Filho mais velho de Eduardo Campos, é candidato a deputado federal Antonio Campos | Neto • Candidato a deputado estadual
Coelho (PE)	Nilo Coelho (1920-1983) • Governou Pernambuco e chegou a presidir o Senado	Fernando Bezerra Coelho | Sobrinho • Foi ministro e é senador

Família e estado	Principal político	Principais familiares na política
		Fernando Filho | Sobrinho-neto • Foi ministro e é deputado federal. Tenta a reeleição Miguel Coelho | Sobrinho-neto • Prefeito de Petrolina (PE) Antonio Coelho | Sobrinho-neto • Candidato a deputado estadual Guilherme Coelho | Sobrinho • Ex-deputado, candidato a 1º suplente de senador
Ferreira Gomes (CE)	Ciro Gomes • Descendente de políticos, foi prefeito, ministro, governador e deputado. Tenta pela terceira vez chegar à Presidência da República	Cid Gomes | Irmão • Ex-governador do Ceará, é candidato ao Senado Ivo Gomes | Irmão • Prefeito de Sobral (CE) Lia Gomes | Irmã • Candidata a deputada estadual Lúcio Gomes | Irmão • Secretário de Infraestrutura do governo do Ceará Tin Gomes | Primo • Deputado estadual, tenta reeleição
Maia (RN)	José Agripino Maia • Filho de político, governou o Rio Grande do Norte, presidiu o DEM e é senador. Vai disputar vaga de deputado federal	Zenaide Maia | Prima • Deputada federal, vai concorrer ao Senado João Maia | Primo • Ex-deputado federal, vai tentar retornar à Câmara Marcia Maia | Prima • Deputada estadual, tenta reeleição Felipe Maia | Filho • Candidato a deputado estadual
Alves (RN)	Aluízio Alves (1921-2006) • Ex-ministro, ex-governador do Rio Grande do Norte	Garibaldi Alves Filho | Sobrinho • Governou o RN e foi ministro. É senador e tenta a reeleição Carlos Eduardo Alves | Sobrinho • Ex-prefeito, candidato ao governo do Estado Henrique Eduardo Alves | Filho • Presidiu a Câmara e foi ministro no governo Temer Walter Alves | Sobrinho-neto • Deputado federal, candidato à reeleição

Família e estado	Principal político	Principais familiares na política
		José Dias | Cunhado • Deputado estadual, candidato à reeleição
Magalhães (BA)	Antonio Carlos Magalhães (1927-2007) • Filho de político, foi deputado, governador, ministro e senador, sendo uma das principais figuras políticas nacionais do Brasil no fim do século 20	ACM Neto | Neto • Prefeito de Salvador Luís Eduardo Magalhães (1955-1998) | Filho • Presidiu a Câmara e era tratado por ACM como seu sucessor político. Morreu de infarto aos 43 anos Paulo Magalhães | Sobrinho • Deputado federal, candidato a reeleição Paulo Magalhães Jr | Sobrinho-neto • Vereador em Salvador
Cunha Lima (PB)	Ronaldo Cunha Lima (1936-2012) • Prefeito, governador da Paraíba, deputado federal e senador	Cássio Cunha Lima | Filho • Ex-governador da Paraíba, é senador e tenta a reeleição Pedro Cunha Lima | Neto • Deputado federal, tenta a reeleição Arthur Cunha Lima | Sobrinho • Deputado estadual, tenta a reeleição Bruno Cunha Lima | Sobrinho • Candidato a deputado federal
Roriz (DF)	Joaquim Roriz Governou o distrito federal, foi ministro e senador. Aos 82 anos, está fora da vida pública	Joaquim Roriz Neto | Neto • Candidato a deputado federal Paulo Roriz | Sobrinho • Candidato a deputado federal Dedé Roriz | Sobrinho • Candidato a deputado distrital
Barros/ Richa/ Requião (PR)	Ricardo Barros, Beto Richa e Roberto Requião • Ex-ministro da Saúde, x-governador e senador. O primeiro é candidato a deputado federal. Os dois últimos, a senador	Cida Borgheti | Mulher de Barros • Governadora, candidata a reeleição Requião Filho | Filho de Requião • Candidato a deputado estadual Marcello Richa | Filho de Richa • Candidato a deputado estadual Maria Victoria | Filha de Barros • Candidata a deputado estadual
Tatto (SP)	Arselino Tatto O primeiro dos irmãos a conseguir um mandato, é vereador desde 1989	Nilto Tatto | Irmão • Deputado federal, candidato à reeleição Enio Tatto | Irmão • Deputado estadual, candidato à reeleição Jilmar Tatto | Irmão • Candidato a senador

Família e estado	Principal político	Principais familiares na política
		Jair Tatto | Irmão • Vereador
Garotinho (RJ)	Anthony Garotinho • Ex-governador do Rio, ex-deputado federal, tenta novamente o governo do Rio	Rosinha Garotinho | Mulher • Ex-governadora do Rio Clarissa Garotinho | Filha • Deputada federal, tenta a reeleição Wladimir Garotinho | Filho • Candidato a deputado federal
Bolsonaro (RJ)	Jair Bolsonaro • Deputado por sete mandatos, lidera hoje as pesquisas de intenção de voto nos cenários sem Lula	Flávio Bolsonaro | Filho • Deputado estadual, é candidato ao Senado Eduardo Bolsonaro | Filho • Deputado federal, é candidato à reeleição Carlos Bolsonaro | Filho • Vereador