Pontifícia Universidade Católica
do Rio de Janeiro

Luiza Frizzo Trugo

Classes de palavras — da Grécia Antiga ao Google: Um estudo motivado pela conversão de tagsets

Dissertação de Mestrado

Dissertação apresentada ao Programa de Pós-Graduação em Estudos da Linguagem da PUC-Rio como requisito parcial para obtenção do título de Mestre em Letras/Estudos da Linguagem.

Orientadora: Profa. Maria Cláudia de Freitas

Rio de Janeiro
Agosto de 2016

Pontifícia Universidade Católica
do Rio de Janeiro

Luiza Frizzo Trugo

Classes de palavras — da Grécia Antiga ao Google: Um estudo motivado pela conversão de tagsets

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Estudos da Linguagem da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

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

Profa. Helena Franco Martins
Departamento de Letras – PUC-Rio

Profa. Sandra Maria Aluísio
USP

Profa. Monah Winograd
Coordenadora Setorial do Centro de Teologia
e Ciências Humanas – PUC-Rio

Rio de Janeiro, 25 de agosto de 2016.

Luiza Frizzo Trugo

Graduou-se em Letras — Bacharelado Bilíngue, na PUC-Rio, em 2013. Concluiu seu mestrado em Estudos da Linguagem, com ênfase nas áreas de Descrição do Português e Processamento de Linguagem Natural, na PUC-Rio, em 2016.

Ficha Catalográfica

Trugo, Luiza Frizzo

Classes de palavras — da Grécia Antiga ao Google : um estudo motivado pela conversão de tagsets / Luiza Frizzo Trugo ; orientadora: Maria Cláudia de Freitas. – 2016.
113 f. : il. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Letras, 2016.
Inclui bibliografia

1. Letras – Teses. 2. Corpus. 3. Linguística computacional. 4. Classes de palavras. 5. Anotação. 6. Particípio. I. Freitas, Maria Cláudia de. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Letras. III. Título.

CDD: 400

Para o meu pai, fonte de inspiração diária, acadêmico de excelência e ser humano raro, que infelizmente partiu muito antes de me ver seguindo seus passos.

Agradecimentos

À minha mãe, por absolutamente tudo, mas especialmente por todos os conselhos dados, por toda a ajuda e pelos incentivos constantes.

À minha orientadora, Cláudia Freitas, pela energia e pela empolgação contagiantes, por topar empreitadas que a maioria não toparia, por me apresentar à fascinante área de PLN e por ser sempre tão presente, rigorosa e bem humorada.

Ao Alexandre Rademaker e ao Fabrício Chalub, da IBM, por toda a ajuda com programação (sem a qual esta pesquisa sequer existiria) e por terem sido sempre tão pacientes e solícitos conosco.

Ao Rafael Rocha, do LEARN (PUC-Rio), por ter sido muitíssimo prestativo ao nos fornecer os dados relativos à acurácia do sistema e à análise de erros.

Àqueles que obsequiosamente participaram do nosso questionário, por terem dedicado seu tempo e esforço classificando um bom número de particípios de difícil classificação.

Ao João Artur, por estar sempre ao meu lado e me fazer rir mesmo nos períodos mais desafiadores.

A Bianca, Juliana e Paula por todo o apoio e pelas longas conversas.

À Tina, por ser a Tina — quem a conhece, entenderá.

Ao *Team PUC*, pois eu definitivamente não teria sobrevivido ao mestrado sem esse incrível grupo de pessoas. Obrigada por tudo.

Aos meus pacientes amigos, por não ressentirem minhas ausências nem esquecerem da minha existência durante esses dois anos.

À PUC, ao CNPq e à FAPERJ, pelos auxílios e bolsas que me permitiram dedicação exclusiva a esta pesquisa.

Resumo

Trugo, Luiza Frizzo; de Freitas, Maria Cláudia. **Classes de palavras — da Grécia Antiga ao Google: um estudo motivado pela conversão de tagsets**. Rio de Janeiro, 2016, 113p. Dissertação de Mestrado – Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

A dissertação "Classes de palavras — da Grécia Antiga ao Google: um estudo motivado pela conversão de tagsets" consiste em um estudo linguístico sobre classes gramaticais. A pesquisa tem como motivação uma tarefa específica da Linguística Computacional: a anotação de classes gramaticais (POS, do inglês “part of speech”). Especificamente, a dissertação relata desafios e opções linguísticas decorrentes da tarefa de alinhamento entre dois tagsets: o tagset utilizado na anotação do corpus Mac-Morpho, um corpus brasileiro de 1.1 milhão de palavras, e o tagset proposto por uma equipe dos laboratórios Google e que vem sendo utilizado no âmbito do projeto Universal Dependencies (UD). A dissertação tem como metodologia a investigação por meio da anotação de grandes corpora e tematiza sobretudo o alinhamento entre as formas participiais. Como resultado, além do estudo e da documentação das opções linguísticas, a presente pesquisa também propiciou um cenário que viabiliza o estudo do impacto de diferentes tagsets em sistemas de Processamento de Linguagem Natural (PLN) e possibilitou a criação e a disponibilização de mais um recurso para a área de processamento de linguagem natural do português: o corpus Mac-Morpho anotado com o tagset e a filosofia de anotação do projeto UD, viabilizando assim estudos futuros sobre o impacto de diferentes tagsets no processamento automático de uma língua.

Palavras-chave

Corpus; linguistica computacional; classes de palavras; anotação; particípio.

Abstract

Trugo, Luiza Frizzo; de Freitas, Maria Cláudia (Advisor). **Part of speech — from Ancient Greece to Google: a study motivated by tagset conversion.** Rio de Janeiro, 2016, 113p. MSc. Dissertation – Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

The present dissertation, "Part of speech — from Ancient Greece to Google: a study motivated by tagset conversion", is a linguistic study regarding gramatical word classes. This research is motivated by a specific task from Computational Linguistics: the annotation of part of speech (POS). Specifically, this dissertation reports the challenges and linguistic options arising from the task of aligning two tagsets: the first used in the annotation of the Mac-Morpho corpus — a Brazilian corpus with 1.1 million words — and the second proposed by Google research lab, which has been used in the context of the Universal Dependencies (UD) project. The present work adopts the annotation of large corpora as methodology and focuses mainly on the alignment of the past participle forms. As a result, in addition to the study and the documentation of the linguistic choices, this research provides a scenario which enables the study of the impact different tagsets have on Natural Language Processing (NLP) systems and presents another Portuguese NLP resource: the Mac-Morpho corpus annotated with project UD's tagset and consistent with its annotation philosophy, thus enabling future studies regarding the impact of different tagsets in the automatic processing of a language.

Keywords

Corpus; computational linguistics; part of speech; annotation; past participle.

Sumário

Lista de quadros

Lista de figuras

Lista de gráficos

1
Introdução

Na área de Processamento de Linguagem Natural (PLN), a anotação (em inglês, *tagging*) de partes do discurso (em inglês, *part-of-speech*, chamado comumente de POS) é uma das formas mais básicas de adicionar informações linguísticas a um corpus, sendo frequentemente essencial e a primeira etapa de processos mais complexos. Tal anotação é realizada por meio da adição de tags, ou etiquetas, às palavras, indicando a que classe gramatical pertencem (Leech, 2005).

A inclusão de anotação de POS em textos é útil para diversos propósitos, como, por exemplo, servir como um pré-processador para níveis de análise mais complexos que se beneficiam das informações refinadas sobre POS (que é um nível mais básico de análise linguística) (Mitkov, 2004). Isso facilita a execução de uma série de tarefas dependentes de dados linguísticos, como a extração de informações, o desempenho de assistentes pessoais inteligentes (como a Siri, da Apple) e, do ponto de vista de estudo de uma língua, a coleta de dados mais refinados para diversos tipos de pesquisas, quando aplicados a grandes corpora. Além disso, de acordo com Manning e Schutze (1999), a anotação de POS possibilita que seja feito no corpus o que é chamado de *partial parsing* (ou parsing superficial), ou seja, possibilita uma anotação sintática automática.

O conjunto das etiquetas utilizadas na anotação de um corpus é chamado de tagset. Existem inúmeros tagsets diferentes, desenvolvidos para diversas línguas – e uma mesma língua pode contar com diferentes tagsets, devenvolvidos por diferentes sistemas ou grupos. Como não há consenso quanto às classes de palavras, há margem para discordância sobre quais categorias são úteis para cada grupo de pesquisa, ou quais são linguisticamente aplicáveis. Há também espaço para interferência de limitações práticas — como o desempenho de um anotador automático —

pois muitas vezes uma distinção que possa parecer teoricamente relevante para um gramático pode, na prática, gerar péssimos resultados em anotadores automáticos (Garside et al, 1997).

Manning (1999) enfatiza que um tagset normalmente incorpora distinções morfológicas da língua para a qual está sendo desenvolvido, não sendo muito simples transpor um tagset de uma língua para outra.

Em Português, temos notícia de cinco tagsets, utilizados pelo etiquetador AnELL[1], pelo etiquetador LAEL[2], pelo projeto Lacio-Web no corpus Mac-Morpho[3], pelo parser PALAVRAS no corpus Bosque[4] e o pelo corpus histórico Tycho Brahe[5].

No entanto, estudos linguísticos que se debrucem sobre diferentes tagsets, comparando-os e/ou investigando seu impacto em tarefas subsequentes são raros e, com relação à língua portuguesa, inexistentes. Uma das exigências para a realização desses estudos é a existência de materiais comparáveis – um mesmo corpus anotado com diferentes tagsets, e corpora distintos anotados com o mesmo tagset —, todos em versões *golden*, isto é, verificados por humanos.

A presente dissertação é parte de um projeto mais amplo que tem como objetivo estudar tagsets e o seu impacto no PLN, o que envolve conversão, revisão e ampliação de material comparável em português. O objetivo desta dissertação é contribuir para a realização deste tipo de estudo, e para tanto estamos criando um cenário que o viabilize: trata-se da conversão do corpus Mac-Morpho (Aluísio et al. 2003), anotado com o tagset do Mac-Morpho, para o tagset desenvolvido por Petrov, Das &

---

1 http://acdc.linguateca.pt/AnELL, acessado em 20/10/2015

2 http://lael.pucsp.br/corpora/etiquetagem/, acessado em 20/10/2015

3 http://nilc.icmc.usp.br/macmorpho/, acessado em 20/10/2015

4 http://visl.sdu.dk/visl/pt, acessado em 20/10/2015

5 http://www.tycho.iel.unicamp.br, acessado em 20/10/2015

McDonald (2011), do Google Research Lab, e atualmente incorporado no projeto Universal Dependencies (UD).

Partindo da ideia de que seria possível estabelecer categorias gramaticais comuns e gerais às línguas, tendo em vista sobretudo a utilização e aproveitamento dos mesmos recursos e ferramentas para o processamento de diferentes línguas, Petrov, Das & McDonald (2011) propõem o que chamam de um tagset "universal", que funcionaria para qualquer língua. Esse tagset, num primeiro momento é composto por 12 categorias, mas atualmente foi ampliado para 14 (ou 17, se contarmos com as categorias para pontuação, símbolos e fragmentos desconhecidos). O desenvolvimento de um tagset "universal", além de possibilitar o trabalho de ferramentas/sistemas multilíngues, também permitiria a comparação de diferentes anotadores treinados em diferentes línguas, mas que compartilham o mesmo tagset.

A ideia de padronização na anotação não é nova. Tentativas como o Eagles[6] existem desde 1996, mas no entanto não vingaram, com cada grupo de pesquisa desenvolvendo seu conjunto de etiquetas. A proposta dos pesquisadores do Google se populariza: (a) devido à força econômica e o impacto social do Google; (b) devido às necessidades reais desse grupo de processar textos de diferentes línguas. Trata-se, portanto, de uma proposta com uma forte base empírica, mas com poucas reflexões linguísticas.

No artigo de 2011, Petrov et al. relatam o alinhamento entre o tagset "universal" e 25 línguas, Português incluído. Nesse caso, o mapeamento teve por base o corpus e o tagset do Bosque (Afonso et al. 2002), com bons resultados: o nível de granularidade das etiquetas do Bosque e do tagset "universal" (que chamaremos daqui em diante de UD, uma vez que integra o projeto Universal Dependencies) é bem próximo, o que facilita a conversão das etiquetas — mas não garante a manutenção da qualidade do

6 http://home.uni-leipzig.de/burr/Verb/htm/LinkedDocuments/annotate.pdf, acessado em 20/10/2015

corpus derivado, como será abordado no capítulo 5. O Bosque, porém, com cerca de 200 mil palavras, das variantes português brasileiro e português de Portugal, é considerado um corpus relativamente pequeno. Para fins de aprendizado de máquina/estatística — o paradigma dominante hoje no PLN (Manning & Schütze, 1999) — tamanho é uma variável relevante, e quanto maior o corpus disponível para treino, mais chances de o aprendizado ser bem sucedido. Além do tamanho, outro aspecto crucial capaz de determinar o potencial de utilização de um corpus anotado é a qualidade da sua anotação pois, no contexto da avaliação de sistemas, ele atua como um "gabarito", indicando os resultados desejados por um sistema. Corpora bons (para treino e para avaliação de resultados) são corpora anotados e/ou revisados por especialistas (humanos). Tanto o Bosque quanto o Mac-Morpho foram revisados, o que garante a sua qualidade.

O tagset do Mac-Morpho é mais granular que o do projeto UD, com diferenças que não são facilmente contornáveis com um alinhamento automático e que fazem da tarefa de alinhamento uma fonte de desafios linguísticos interessantes para a descrição da língua portuguesa. No âmbito desta dissertação, dois grandes desafios linguísticos vinculam-se às classes "Particípio" e "Palavras Denotativas", presentes apenas no tagset do Mac-Morpho. Tais classes são reconhecidamente desafiadoras na literatura, e sua classificação não é consensual. A literatura linguística mostra algumas soluções para essa questão (detalhadas no capítulo 3 desta dissertação), porém estão longe de serem satisfatórias.

Todas as decisões linguísticas tomadas durante o mapeamento dos tagsets estão minuciosamente documentadas e discutidas nesta dissertação, dando origem a um trabalho detalhado de descrição linguística com base em corpus e motivado pelo PLN. Adicionalmente, o alinhamento entre os tagsets, que tem como resultado o corpus Mac-Morpho com uma segunda anotação, viabiliza um estudo sobre o impacto de diferentes tagsets em uma tarefa de anotação – trabalho de grande relevância e que até bem pouco tempo não havia como ser feito.

1.1

Objetivos

Os objetivos gerais desta dissertação são: (a) a construção de recursos linguísticos que viabilizem investigação sobre o impacto de diferentes tagsets no PLN da língua portuguesa. Esse recurso consiste em dois datasets distintos de português anotados com dois tagsets diferentes (tagset UD e um segundo, híbrido, que contém as classes presentes no UD e, adicionalmente, a etiqueta "particípio"); (b) investigar a adequação, para a língua portuguesa, de um tagset planejado para dar conta de diversas línguas – um tagset chamado de "universal"; ou seja, averiguar se o desempenho de um sistema é melhor e mais eficiente em um corpus em português quando se utiliza um tagset pensado especificamente para o português ou um criado com o objetivo de ser universal.

Os objetivos específicos são (i) investigar as decisões linguísticas tomadas na anotação dos corpora originais, (ii) discutir e documentar as decisões linguísticas tomadas ao longo do alinhamento, contribuindo assim para a descrição do português e (iii) realizar um estudo com base em corpus, sobre a classe dos particípios, capaz de oferecer subsídios para as decisões linguísticas da anotação.

1.2

Motivação e Justificativa

Tagsets são metalinguagens linguísticas apropriadas para o PLN. Apesar de seu caráter intrinsecamente linguístico, não se tem notícias de estudos linguísticos que tematizem tagsets quanto ao impacto no desempenho de sistemas de PLN. Uma possível razão para essa lacuna é justamente a carência de recursos que permitam esse tipo de estudo – tanto do ponto de vista linguístico, quanto do ponto de vista dos algoritmos que realizam a tarefa de POS.

Com relação ao tagset UD, desde 2011 o corpus Bosque está anotado com esse conjunto de etiquetas (Patrov et al., 2011). No entanto, essa anotação foi feita inteiramente de forma automática, por meio do alinhamento das etiquetas (que compartilhavam diversas características, dentre elas o nível de granularidade, o que facilita o alinhamento)[7]. Além disso, embora o Bosque tenha amplo uso na comunidade internacional, devido sobretudo à sua participação no CONLL de 2006[8], o corpus Mac-Morpho é outra grande referência para POS em língua portuguesa, sobretudo no âmbito nacional, devido à sua extensão, 1,1 milhão de palavras. Outra característica positiva do Mac-Morpho é a sua documentação linguística detalhada, fundamental na etapa de alinhamento. Quanto ao conteúdo, ambos os materiais são parecidos, pois tratam-se de textos jornalísticos publicados na Folha de São Paulo (Mac-Morpho e Bosque) e no jornal Público (Bosque).

O alinhamento dos tagsets Mac-Morpho–UD, no entanto, é bem menos óbvio que o alinhamento Bosque-UD. Isso ocorre porque o tagset do Mac-Morpho é mais granular, e, como já mencionado, duas classes, especialmente, despontam como problemáticas para a conversão automática: "Particípio" (PCP) e "Palavras denotativas" (PDEN). Não há, no tagset UD, nenhuma categoria diretamente equivalente a estas.

Assim, a presente pesquisa, portanto, dialoga diretamente com as áreas de Descrição Linguística e Processamento de Linguagem Natural (PLN). Trata-se de uma contribuição para a descrição do português motivada por uma tarefa de PLN — no caso, converter um corpus anotado para um segundo tagset.

A dissertação está organizada da seguinte maneira: no capítulo 2, tratamos do enquadramento teórico que norteou este trabalho; no capítulo 3,

---

[7] O alinhamento está disponível em http://universaldependencies.github.io/docs/tagset-conversion/pt-conll-uposf.html, acessado em 20/10/2015

[8] O CONLL é uma compteição anual entre sistemas. O Bosque foi usado no CoNLL-X, em 2006, cuja tarefa era análise sintática dependencial multilingüe.

abordamos a metalinguagem gramatical e a história das classes de palavras, apresentando maior foco nos particípios e nas chamadas palavras denotativas; no capítulo 4, apresentamos a metodologia empregada nesta pesquisa e os corpora/tagsets Mac-Morpho e UD; no capítulo 5, propomos uma conversão do tagset Mac-Morpho para o UD, discorrendo sobre as decisões linguísticas tomadas no processo de alinhamento; no capítulo 6, discutimos os impactos da conversão e implicações para sistemas de PLN, utilizando como base a análise de erros de um anotador automático. Por fim, tecemos algumas considerações finais e apresentamos possibilidades de aprimoramento e expansão do trabalho conduzido.

2
Enquadramento teórico

Este trabalho assume uma perspectiva não-logocêntrica no diálogo com o PLN e com a Descrição do português, como anunciado em Freitas (2007; 2013) e utilizado em Freitas (2016). Uma visão que assume, com Auroux (1992), que o saber linguístico é um produto historicamente constituído, localizado em um tempo e em um espaço. Assume-se que a descrição de uma língua será sempre parcial e motivada por interesses, sendo as fontes de dados para essa descrição os grandes corpora.

Um corpus é uma coleção de objetos linguísticos, classificada, finita e concreta, podendo ou não ser anotada, que pode representar o falante comum e que é criada com o intuito de servir como "utensílio para estudar a língua (ou literatura ou cultura)" (Santos, 2008). Santos dá ainda alguns exemplos do que exatamente podem ser tais objetos linguísticos: "textos, frases, palavras, entrevistas, erros ortográficos, entradas de dicionário, citações, pareceres jurídicos, filmes, imagens com legendas, traduções..." etc.

Assim, um corpus fornece as ferramentas necessárias para que o linguista observe a língua como ela é utilizada, possibilitando uma postura empírica diante dela. Pode-se então utilizar os dados concretos e observáveis fornecidos pelo corpus para elaborar novas categorias ou teorias capazes de descrever a língua, ao invés de partir da intuição linguística pessoal de um falante para elaborar uma teoria que pode ou não ser de fato aplicável a língua como ela é utilizada. Conforme Sampson (2001) afirma, "so long as science strives to find itself on interpersonally observable data, it can always move forward through critical dialogue among the community of researchers", ou seja, dados observáveis são a melhor fonte para se elaborar teorias, pois estão disponíveis para todos, de forma que um diálogo crítico torna-se possível e, através dele, as teorias tendem a ser aprimoradas.

Assim, a relação torna-se mais proveitosa quando se parte do corpus para a teoria, e não o contrário, já que é possível testar a teoria amplamente através dos dados.

No diálogo com os estudos com base em corpus, assumimos uma visão segundo a qual o linguista é não é mais o

> "falante-ouvinte ideal em uma comunidade de fala completamente homogênea, que conhece a sua língua perfeitamente, mas sim de um falante-ouvinte comum em uma comunidade heterogênea, que conhece a sua língua apenas parcialmente e, de maneira ativa, busca acesso ao conhecimento de outros. Nossas afirmações têm autoridade não devido aos superpoderes da introspecção, mas do exame de grandes conjuntos de dados autênticos (...)" (Beaugrande, 2002)

Por fim, assumimos uma perspectiva que entende a anotação linguística não apenas como uma atividade do PLN, mas como uma forma de investigação linguística, como defendido em Sampson (2001), Archer (2012); Santos et al (2015) e Freitas (2015).

2.1
Descrição e PLN

A área de descrição do Português (ou de qualquer outra língua) e a área do PLN encontram-se e beneficiam-se mutuamente na pesquisa com corpus, de acordo com Freitas (2013), podendo se relacionar basicamente de duas formas. Uma destas seria quando a descrição da língua é motivada pela aplicação, ou seja, pela necessidade de resolver problemas e tarefas de sistemas que manipulam a língua, o que acaba por expandir nossa compreensão sobre a própria língua.

Outra maneira seria quando a descrição se utiliza de ferramentas da Linguística Computacional – como concordanciadores e analisadores

morfossintáticos – para estudar algum aspecto da língua, contribuindo para o entendimento de certo fenômeno linguístico. Com relação à língua portuguesa, Garrão et al.(2008 apud Freitas 2013) tratam da identificação de sintagmas preposicionais em corpus; Oliveira (2006 apud Freitas 2013), da caracterização de substantivos vazios, e Freitas et al. (2012 apud Freitas 2013) realizam um estudo sobre as cores em língua portuguesa.

Para o PLN, um corpus é um recurso linguístico que, ao ser anotado com informações linguísticas de diversas naturezas, torna-se extremamente útil para a execução de tarefas variadas (como extração de informação). Afinal, como argumenta Aluísio (2011), (i) métodos computacionais são relevantes, mas o investimento em bons corpora é mais importante ainda, pois "bons corpora anotados duram décadas", enquanto "métodos são substituidos por novos métodos mais rapidamente" e ii) um projeto de corpus, quando mal conduzido, tem a capacidade de prejudicar anos de pesquisa. Já do ponto de vista linguístico, o corpus é um meio promissor de estudar empiricamente a língua, sendo uma fonte excelente para a descrição linguística (Sampson, 2001; , Beaugrande, 2002; Archer, 2012; Santos, 2015).

Como um corpus apresenta porções reais da língua como é usada, não é possível fugir das "irregularidades" da língua (frequentemente deixadas de fora do escopo dos estudos linguísticos justamente por fornecerem poucos insights sobre as regularidades/sobre a estrutura) durante o processo de anotação. Assim, é preciso lidar com casos difíceis, periféricos e/ou pouco descritos, o que acentua ainda mais a característica da anotação de ser um processo de análise e pesquisa.

A anotação de um corpus pode ocorrer por meio da delimitação de segmentos de um texto (que podem ser palavras, expressões etc) para a atribuição de etiquetas (categorias) a esses segmentos, ou para o estabelecimento de relações entre os segmentos (como por exemplo na anotação de relações semânticas ou de correferência). As etiquetas são

definidas de acordo com o objetivo da anotação, de forma que é necessário decidir a forma de lidar com o problema em questão para que se possa definir o conjunto de etiquetas que será utilizado na anotação.

Muitas vezes, o processo de anotação é visto somente como uma etapa anterior necessária à execução de outras etapas ou tarefas futuras. Porém, autores como Santos (2014), Sampson (2001), Archer (2012) e Freitas (2015) defendem que a anotação de corpora não é apenas uma atividade mecânica com o objetivo de fornecer informações para sistemas de processamento automático de língua, mas também um meio de investigação e estudo empírico da língua. Para eles, corpora anotados são a matéria-prima para uma parcela significativa dos estudos linguísticos modernos.

Ao anotar, é necessário categorizar e classificar um fenômeno, o que é uma forma de estabilização, e o próprio processo de anotação faz com que o anotador seja confrontado pelos limites dessas estabilizações, que podem parecer teoricamente claros, mas na prática revelam-se muito mais tênues e incertos. Assim, a anotação sempre refletirá interpretações e posicionamentos que o pesquisador teve de tomar ao longo do processo de anotação.

Para Sampson (2001), especificamente, o processo de anotação é o "substituto moderno" da forma tradicional de se fazer pesquisa, e a anotação é assim um processo de interpretação, classificação e formalização de um determinado fenômeno, havendo sempre a possibilidade de discordância e interpretações distintas por parte de diferentes pesquisadores/anotadores[9]. Mesmo a anotação de informações consideradas objetivas, como as classes de palavras, são fruto de uma perspectiva teórica (nesse caso, a das

[9] De fato, um dos principais objetivos de Santos et al (2015) com a Gramateca é, por meio da anotação, contribuir com a metodologia científica nos estudos linguísticos, permitindo a repetição de experiências (propriedade essencial da metodologia científica) e partilhando diferenças de interpretação relativas a um mesmo material. Os autores defendem que o compartilhamento de material classificado linguisticamente tem o potencial para servir de base para mais estudos sobre a gramática da língua portuguesa.

gramáticas tradicionais), como lembra Freitas (2015). E mesmo nesses casos é possível haver discordância entre os anotadores — como é o caso das formas participiais, que abordaremos em detalhes nos próximos capítulos desta dissertação.

Um ponto interessante ressaltado ainda em Santos et al (2015) é que, nas ciências humanas, diferentemente das ciências exatas ou biológicas, não se pode esperar que as mesma experiências gerem os mesmos resultados — muito pelo contrário, há grandes possibilidades de que diferentes pesquisadores, ao repetir determinada experiência, tenham interpretações distintas —, mas que essa característica costuma ser criticada ao invés de ser aceita como parte de um tipo diferente de ciência – uma diferença entre ciências humanas e exatas. Interpretações diferentes não devem ser vistas como desviantes ou anômalas, mas sim como alternativas que enriquecem as discussões e os estudos linguísticos (o que não significa que todas as alternativas sejam, sempre, igualmente válidas). Assim, conforme os autores, não há sentido em comparar ciências humanas a ciências exatas/ biológicas ou em tentar igualá-las, pois são conhecimentos de naturezas distintas.

Em Santos et al (2015), os autores levantam uma interessante questão: quando se para de pensar no cerne da tarefa de anotação como a construção de um recurso e começa-se a pensar que a anotação pode ser uma forma de investigar fenômenos linguísticos, torna-se interessante explicitar as discordâncias, pois estas são justamente o reflexo de diferentes interpretações a respeito de um mesmo fenômeno. Os autores enfatizam que são justamente as divergências que "alimentam" a pesquisa, sendo extremamente relevantes no estudo de uma língua.

Santos (2014) acredita que estamos em um momento propício para tornar possível a realização de estudos gramaticais de larga escala estatisticamente informados e ainda possibilitar a consulta ao material

utilizado, sendo este público. Ela defende que tanto o PLN quanto a linguística têm muito o que contribuir para a gramática do Português.

No contexto do PLN, Sampson (2012) afirma que, na literatura de processamento de linguagem natural, não é comum a discussão aberta sobre as análises subjacentes às coleções "golden" (que funcionam como gabaritos), ou seja, não é comum a discussão sobre o tipo de análise lingüística que, em última análise, é alvo dos sistemas. O autor sugere que o fato de análises gramaticais serem estudadas exaustivamente durante os anos de colégio dá aos pesquisadores a falsa sensação de que esses conhecimentos são dominados por todos e de forma igual, não sendo assim necessário explicitar tais análises. Essa noção, porém, está muito equivocada, e o exemplo fornecido por Sampson de um episódio ocorrido na conferência "Association of Computational Linguistics" de 1991, na universidade de Berkeley, ilustra o quão irreal é essa expectativa: durante um workshop, representantes de 9 instituições receberam uma série de frases retiradas de corpora e deveriam indicar a análise que suas instituições ou seus laboratórios têm como alvo para aquela determinada frase. Para a surpresa dos participantes, as semelhanças entre as análises foram pouquíssimas; até mesmo a divisão de constituintes foi completamente diferente. Sampson utiliza esse ponto para argumentar que deveria haver mais incentivos para que os pesquisadores discutissem pública e abertamente as análises gramaticais e sintáticas que utilizam em seus corpora, pois há um foco desproporcionalmente maior nos sistemas utilizados e costuma-se ignorar o conhecimento linguístico subjacente. O autor aponta ainda que muitos preconceitos já foram quebrados em relação à área de PLN; hoje em dia, a compilação de corpora é vista como algo importante e interessante em que investir, sendo muito mais valorizada do que há algumas décadas. Além disso, afirma que a maioria dos linguistas empíricos já vêem o sentido e a utilidade de trabalhar com corpus anotado, por exemplo. Porém, até hoje ainda não se parece ter entendido a

importância de dedicar tempo e verbas no refinamento no estudo e na discussão dos esquemas de anotação.

Como já foi dito, Sampson (2001) ressalta que nos estudos linguísticos estruturalistas, o foco de interesse está naquilo que é "ordenado" e "homogêneo", frequentemente ignorando a língua em uso (por esta ser "caótica") e tomando por base exemplos inventados e perfeitamente controlados. O PLN, porém, por precisar se basear na língua como de fato é utilizada, não pode escolher as estruturas com as quais deseja trabalhar e deixar outras de lado – é necessário lidar com os mais variados aspectos e estrutura, mesmo aqueles considerados menores ou periféricos (como as palavras denotativas). Ele ressalta que nessa área o que importa é resolver os problemas a medida que vão surgindo, desenvolvendo formas confiáveis, práticas e consistentes de registrar e analisar a variedade de dados que se encontram em um corpus.

Outra dimensão do trabalho com corpus, na perspectiva assumida aqui, diz respeito à documentação, sobretudo quanto às opções linguísticas. Como já mencionado, é razoável que haja diversidade de interpretações linguísticas e, frequentemente, mais de uma alternativa pode ser válida, de forma que não há soluções únicas. Assim, o registro das soluções adotadas para tratar determinado fenômeno torna-se essencial. Nesse contexto, o próprio ato de documentar já é em si uma descrição da língua, pois nesse processo são capturadas regularidades, exceções e casos híbridos ou não consensuais.

As ideias e posicionamentos apresentados aqui guiaram a elaboração da presente dissertação, que não deixa de ser — dentre outras coisas — uma pesquisa sobre anotação.

3
Sobre a metalinguagem gramatical

3.1
A história das classes ao longo dos séculos

Primeiramente, consideramos relevante reiterar que, ao longo desta dissertação, assumimos a posição de que classes gramaticais são objetos historicamente construídos. Para tanto, nos apoiamos sobretudo na leitura de Auroux (1992), Bagno (2011) e Moura Neves (2005). Apesar de crianças no mundo inteiro estudarem gramática desde cedo na escola como se fosse algo dado, uma verdade absoluta, os saberes sobre a linguagem foram sendo construídos por pensadores ao longo de séculos. Auroux (1992) ressalta que só foi possível desenvolver os estudos linguísticos graças ao surgimento da metalinguagem gramatical, e que apenas por meio desta é possível construir conhecimento sobre a linguagem.

De acordo com Auroux (1992), a gramática é a área de estudo que apresenta o "vocabulário próprio" mais antigo e estável. Quando fala em "vocabulário próprio", o autor refere-se às metalinguagens desenvolvidas para estudar a língua, em especial às classes gramaticais. O que se sabe a respeito das classes gramaticais é que os modelos mais similares aos que utilizamos atualmente parecem ter surgido primeiramente na Grécia Antiga, quando começou a haver a necessidade de falar sobre a língua e de descrevê-la (Moura Neves, 2005). Essa necessidade partiu dos filósofos da época, que propunham questões e buscavam respostas a respeito da existência, da natureza, da lógica, e o interesse sobre a linguagem estava atrelado a esses outros campos. É interessante destacar que as classes existentes hoje em dia para o português e presentes na Nomenclatura Gramatical Brasileira (NGB)[10] não são tão diferentes assim de algumas que já haviam sido propostas por volta de 100 AC, conforme veremos mais adiante.

[10] Documento disponível em https://docs.ufpr.br/~borges/publicacoes/notaveis/NGB.pdf

Como os filósofos procuravam organizar e sistematizar a língua de maneira a responder às suas próprias indagações, não havia (e nem há) um único modelo de "partes do discurso". Por exemplo, Platão propunha apenas duas partes do discurso: *ónoma* e *rhema*, que podem ser traduzidos aproximadamente como sujeito e predicado (ou nome e verbo) – estruturas básicas das proposições lógicas, como indica Auroux (1992). Uma curiosidade apontada por Rosa (2002) é que, nessa classificação, os adjetivos estavam incluídos na parte dos verbos – uma vez que adjetivos, assim como verbos, predicam –, retratando uma visão bastante diferente da que temos hoje em dia e evidenciando a relevância do critério interesse na delimitação da classes ou partes do discurso, que não são tão claras e sólidas como inicialmente parecem, e que a fronteira entre uma classe e outra já foi (e ainda é) tênue. Além disso, vale lembrar que um dos grandes desafios desta dissertação, distribuir as formas participiais entre as classes de verbos e adjetivos, provavelmente não seria um problema para Platão. Isso ocorria porque Platão considerava que a principal função da sentença era a predicação, de forma que a *rhema* sempre atribuiria atributos/ações/ qualidades/estados à *ónoma* (Bagno, 2011: 407)

Aristóteles seguia uma estratégia parecida, mas apresentando três classes ao invés de duas: (i)nomes, (ii)verbos e (iii)conectivos, que incluíam preposições, conjunções e pronomes. Villalva & Silvestre (2014) ressaltam que a separação de unidades lexicais em classes de palavras surge de uma visão aristotélica de linguagem, em que as palavras são distribuídas em classes que lhes atribuem propriedades comuns.

Moura Neves (2005) afirma que o registro gramatical mais antigo de classes de palavras (chamadas então de "partes do discurso") similares aos modelos atuais ao qual se teve acesso é o de Dionísio da Trácia (170 a.C. — 90 a.C.) —,que inclusive apresenta muitas semelhanças com as classes utilizadas para Português atualmente. Ele apresentava a seguinte lista de partes do discurso (também encontrada nos registros de outros filósofos gregos posteriores): nome (englobando o que hoje chamamos de

substantivos e adjetivos), pronome, verbo, advérbio, particípio, conjunção (conjunções e preposições) e artigo (contendo pronomes e artigos). Esse sistema perdurou praticamente inalterado por quase dezessete séculos, desde sua criação até a Baixa Idade Média (Biderman, 2001). A tabela abaixo, retirada de Bagno (2011: 417) ilustra os diferentes modelos de partes do discurso aos quais foi possível obter acesso (ou seja, que não foram destruídos, perdidos e/ou esquecidos ao longo da história).

<table>
<tr><th>Heráclito</th><th>Platão</th><th>Aristóteles</th><th>estoicos</th></tr>
<tr><td rowspan="6">1. lógos ("linguagem, razão, natureza, cosmo")</td><td rowspan="3">1. ónoma ("nome"): substantivo/ sujeito</td><td rowspan="2">1. ónoma ("nome")</td><td>1. ónoma ("nome"): substantivos e adjetivos</td></tr>
<tr><td>2. rhêma ("verbo")</td></tr>
<tr><td rowspan="3">2. rhêma("verbo")</td><td>3. sýndesmos ("conjunção"): conjunções e preposições</td></tr>
<tr><td rowspan="3">2. rhêma ("verbo"): verbo/ predicado</td><td>4. árthron ("artigo"): pronomes retos e pessoais; artigos definidos e demais pronomes</td></tr>
<tr><td rowspan="2">3. sýndesmos ("conectivo"): preposições, conjunções, pronomes</td><td>5. metokhé ("particípio")</td></tr>
<tr><td>6. epírrhema ("advérbio")</td></tr>
</table>

Figura 1: Diferentes modelos de partes do discurso (Bagno, 2011:417)

A partir dessas partes do discurso, desenvolveu-se a tradição da gramática latina, da qual a língua portuguesa é herdeira. Afinal, a gramática do Português não foi criada a partir do zero especificamente para esta língua, o que ocorreu foi uma adaptação a partir de modelos já existentes (originalmente criados para descrever o grego antigo), que não necessariamente são totalmente adequados para o português atual nem para todas as questões que vão surgindo em relação à língua. Assim, não é surpresa alguma afirmar que a tradição gramatical do Português foi sendo construída e modificada durante vários séculos, processo que ainda ocorre e continuará ocorrendo enquanto o Português existir como língua viva. Isso fez com que essa terminologia acumulasse usos heterogêneos e apresentasse metalinguagens de diferentes períodos (Villalva & Silvestre, 2014).

Dessa forma, é importante manter em mente que as classes de palavras que aprendemos na escola podem, a princípio, parecer absolutamente estáveis e incontestáveis, mas tais esquemas classificatórios não são naturais nem homogêneos. Além disso, as fronteiras entre classes muitas vezes são tênues e de difícil delimitação — e a flutuação entre substantivos e adjetivos ilustra bem a questão, como problematizado em Perini (1997) e Basílio (2008), dentre outros.

Isto porque as classes são uma forma de distribuir as palavras da língua em grupos e, como em qualquer classificação, tem seus critérios definidores escolhidos em função dos diferentes interesses na classificação. Os interesses podem privilegiar aspectos formais, de sentido ou discursivos, por exemplo. Diferentes escolas, ao longo da história, se interessaram pelas palavras, privilegiando diferentes aspectos ao longo do tempo. Os estoicos, por exemplo, privilegiavam a dimensão formal, daí a sua grande divisão entre palavras variáveis e invariáveis. Platão se interessava pela natureza das proposições lógicas, compostas por sujeito e predicado.

Houve, ao longo da história, tentativas de formular uma gramática mais específica do Português, fugindo das nomenclaturas e definições de bases latinas. Esses esforços, porém, nunca foram muito apreciados ou incentivados pela maioria dos gramáticos de maior renome e não vingaram (Villalva e Silvestre, 2014). Inclusive, houve momentos em que o oposto foi feito: durante e após o Renascimento, por exemplo, procurou-se voltar o máximo possível às tradições grego-latinas, como forma de "valorizar" o Português, e foi também nessa época que se passou a encarar essas classificações como interlinguísticas e universais, como se pode averiguar nos trabalhos de Escalígero (1540), Francisco Sánchez (1587) e Amaro Reboredo (1619) (todos apud Villalva e Silvestre, 2014).

Como enfatiza Bagno (2011), é importante conhecer o passado da gramática para que não nos esqueçamos das suas origens. Essa "amnésia da gênese" (expressão de Pierre Bourdieu citada por Bagno) pode levar à crença de que a gramática sempre existiu, de que é algo dado, uma verdade

absoluta, o que definitivamente não é o caso. As classes gramaticais não são naturais, não são como o ar, a água, pedras, plantas ou outros elementos cuja existência é bastante óbvia: são, na verdade, tentativas artificiais e arbitrárias de categorização e sistematização do fenômeno linguístico, elaboradas por seres humanos em determinado contexto histórico e social. Bagno inclusive compara a categorização das palavras com a classificação das estrelas em constelações; as estrelas de fato existem, mas constelações não passam de criações humanas para organizá-las e dar-lhes sentido. Dessa forma, pessoas em diferentes épocas e contextos socio-culturais, com interesses distintos, elaboram critérios de categorização diferentes. Como as classes gramaticais também são fruto do contexto em que foram elaboradas e de determinadas decisões teórias, sempre de certa forma arbitrárias, estão sujeitas a reformulações, objeções e críticas.

O reconhecimento de que as classes de palavras são sempre fruto de escolhas humanas, guiadas por diferentes critérios/interesses ao longo do tempo, abre espaço para questionamento relacionado às classes atuais, no âmbito dos estudos linguísticos e do PLN, ao mesmo tempo em que oferece uma justificativa linguisticamente motivada para o investimento na discussão e proposta de outras classes/classificações. Igualmente, deve-se entender que o fato de a classificação gramatical apresentada pelas gramáticas tradicionais do Português atualmente ser a mais utilizada, sendo "mais ou menos consensual, mais ou menos adequada" (Villalva & Silvestre, 2014) não a impede de apresentar pontos nebulosos. As seções a seguir demonstrarão alguns desses desafios classificatórios vinculados aos objetivos desta dissertação.

3.2
Particípios em foco

Considerando que o nosso maior desafio no desenvolvimento deste trabalho foi converter a etiqueta "particípio" para as etiquetas "verbo" ou

"adjetivo" (cf. capítulo 5), torna-se essencial para a compreensão dessa dificuldade que relatemos os pontos mais relevantes a respeito dessa categoria e de sua história.

3.2.1
Análises linguísticas

Desde os primeiros filósofos estoicos (301 a.C.), muitos estudiosos consideravam particípios "nomes verbais", "verbos com casos", "verbos participiais", dentre outros (Rosa, 2002), o que enfatiza suas características mais verbais — apesar de os particípios também se flexionarem em caso, gênero e número, como os nomes. Pode-se ter uma noção mais clara do comportamento das classes na tabela abaixo, retirada de Auroux (1992:106):

	Flexionadas					Não-flexionada		
Denominação	Nome	Verbo	Particípio	Artigo	Pronome	Preposição	Advérbio	Conjunção
Definição	O nome é uma parte da frase casual que designa um corpo ou uma ação e que se emprega como valor comum ou particular.	O verbo é uma palavra não casual que admite tempo, pessoas e número, e que exprime o ativo e o passivo.	O particípio é uma palavra que participa da propriedade dos verbos e da dos nomes.	O artigo é uma parte da frase anteposta (ou posposta) à flexão dos nomes, há o prepositivo ho e pospositivo hos.	O pronome é uma palavra empregada no lugar de um nome e que indica as pessoas definidas.	A preposição é uma palavra que se antepõe a todas as partes da frase em composição e em construção.	O advérbio é uma parte da frase não-flexionada, dita do verbo ou aplicada ao verbo.	A conjunção é uma palavra que conjuga o pensamento pela ordenação e que revela o implícito da expressão.
Acidentes	5	8	7	3	6	0	(2)	0
Gênero (3)	+	−	+	+	+			
Espécie (2)[1]	+	+	+	−	+		(+)	
Figura (3)[2]	+	+	+	−	+		(+)	
Número (3)	+	+	+	+	+			
Caso (5)	+	−	+	+	+			
Modo (5)	−	+	−	−	−			
Diátese (3)	−	+	+	−	−			
Pessoa (3)	−	+	−	−	+			
Tempo (3)[3]	−	+	+	−	−			
Conjunção (13)[4]	−	+	−	−	−			
Sentido Específico[5]	Designa um existente qualificado	Ação (Pragma)[6] Diátese	(Cf. Verbo) substitui um verbo com construção de um nome	Anáfora acompanha um nome	Deixis = referência definida a um existente visível (+ anáfora) substitui o nome	(Relação)	(cf. Objetivo) predica o verbo	ligação interfrástica

Figura 2: "Partes do discurso e categorias gramaticais na tekhnê de Dionísio da Trácia" (Auroux, 1992:106)

Bagno (2011) relata que, quando foi feita a tradução das classes de palavras do grego para o latim, o particípio (*participium*) recebeu esse nome justamente por "participar" tanto da classe dos nomes quanto da dos verbos. Considerava-se que particípios "participavam" da classe dos nomes devido à ausência de Modo e à presença de Caso e Gênero, e da classe dos verbos devido ao fato de apresentarem Tempo e flexões derivadas do verbo. É importante ressaltar que, tanto em grego quanto em latim, o particípio poderia se manifestar nos tempos passado, presente e futuro. Em Português, existe apenas o particípio passado, mas podemos encontrar vestígios do particípio presente em formas como "estudante" (que estuda) e do particípio futuro em formas como "duradouro" (que há de durar).

Atualmente, o particípio é formalmente considerado pela NGB uma forma verbal — encontrando-se entre as chamadas formas nominais do verbo, um rótulo que evidencia seu caráter híbrido, mas privilegia a dimensão verbal — e assim é tratado na maioria das gramáticas. Sintaticamente, porém, particípios podem exercer funções normalmente desempenhadas por adjetivos ou até substantivos (exemplos: "Dos 30 milhões de domicílios com computador no país, só 300 mil usam consistentemente os serviços bancários **informatizados**" e "[…] irão hoje fazer a doação os **eleitos** do PSDB e seus adversários do PDT", respectivamente), além de se flexionarem em gênero e número e terem alguns de seus usos já tão cristalizados que é difícil não considerá-los adjetivos:

(1) "um mix de rock **pesado** com acordes de composição eruditas",

(2) "o público encontra até revistas **especializadas** voltadas especificamente para a família",

(3) "neste mundo **globalizado**, onde nada se faz sem discurso […]"

(4) "é muito **complicad**o ele deixar o governo em abril"

(5) “Decisões **isoladas** desse tipo não vão equacionar a problemática ambiental”[11]

(6) “O vento **enfurecido** açoitava a rancharia”[12]

No entanto, esta interpretação não é unânime: o exemplo (5) é retirado de Moura Neves (2011), que o trata como particípio (e, portanto, verbo), e (6), retirado de Cunha & Cintra, é um exemplo de “tempo composto sem auxiliar” que, por não expressar relação temporal, “se confunde com o adjetivo” (Cunha & Cintra 2001:510).

Ou seja, há formas participiais que são classificadas com unanimidade como verbos, outras são classificadas sem maiores polêmicas como adjetivos e aceitas como tal, e ainda há um enorme número de casos em que os particípios apresentam propriedades sintáticas tanto de adjetivos como de verbos, impossibilitando a identificação plena e indubitável com uma dessas categorias. De forma bastante resumida, acreditamos que a grande dificuldade na classificação dos particípios se dá devido ao fato de atualmente, em português, não formarem um grupo uniforme.

Particípios integram estruturas passivas e tempos compostos, e é nesse contexto – verbal – que costumam ser tratados nas gramáticas atualmente, e de forma bastante superficial. Em geral, os casos mais controversos dizem respeito às estruturas em que não há o verbo auxiliar – pista formal facilitadora, justamente por indicar a presença de construções passivas ou de tempos compostos.

Na gramática de Cunha & Cintra (2001), os particípios encontram-se em uma subseção de um capítulo sobre verbos e destaca-se sua função de exprimir o aspecto conclusivo de um processo verbal. Sobre os particípios em tempos compostos, afirma-se que (i) com os auxiliares “ter” e “haver”, o particípio forma “tempos compostos da voz ativa”, como no exemplo “Temos **estudado** muito” (2001:508); (ii) com o auxiliar “ser”, forma “tempos de voz passiva de ação”, como no exemplo “A carta foi **escrita** por

[11] Exemplo retirado de Moura Neves (2005)

[12] Exemplo retirado de Cunha & Cintra (2001)

mim" (2001: 508) e (iii) com o auxiliar "estar", forma "tempos da voz passiva de estado", como exemplo "Estamos **impressionados** com a situação" (2001: 508). Quando o particípio não é precedido de auxiliar, Cunha & Cintra afirmam que sua função é fundamentalmente exprimir o estado resultante de uma ação acabada (exemplo: "**Achada** a solução do problema, não mais torturou a cabeça" — 2001:508), classificando o valor do particípio de acordo com a transitividade do verbo: particípios de verbos transitivos, de acordo com os autores, sempre apresentariam valor passivo (exemplo: "**Lidas** uma e outra, procedeu-se às assinaturas" — 2001:509) e particípios de verbos intransitivos teriam "quase sempre" valor ativo (exemplo: "**Chegado** aos pés, olhava-me para cima" — 2001:509). Por fim, afirmam que "quando o particípio exprime apenas o estado, sem estabelecer nenhuma relação temporal, ele se confunde com o adjetivo", como no exemplo, já mencionado, "O vento **enfurecido** açoitava a rancharia" (2001:510). Apesar da aparente clareza dos critérios, é evidente o quanto são dependentes de interpretação, e é exatamente este o ponto escorregadio. Nas frases "Durante a festa, apareceu um cantor **vestido** de Elvis Presley" (Mac-Morpho) e "É uma história **centrada** nos seus medos, «nas suas bravatas» (como diz a apresentação), nas suas certezas «e, sobretudo, nas suas dúvidas»" (Mac-Morpho), por exemplo, estamos diante de valor ativo? Ou apenas de estado?

Para Ilari & Basso (2014:232) no exemplo abaixo (retirado do NURC) há o que chamam de "falsa passiva analítica", por não haver particípio passado, mas sim um adjetivo. No entanto, a frase exemplo possui duas estruturas com forma participial, e apenas uma é mencionada como "falsa passiva" (restrita), não havendo comentários para "publicada":

(7) "a atuação dos professores franceses, sobretudo dos mais jovens [...] que não tinham ainda obra […] **publicada**, como era o caso de Jean Moguet e de Claude Lévi-Strauss… era **restrita** e se exercia sobretudo através dos cursos, não atingindo grande público"

Porém, os autores não nos informam a estratégia seguida para identificar as palavras destacadas como adjetivo —- e não como particípios — e, infelizmente, essa é a única menção à estrutura. Esse exemplo assemelha-se aos exemplos de Cunha & Cintra, para os quais não nos parece muito clara a classe que deve ser atribuída, uma vez que "o PCP se confunde com o ADJ". Como, afinal, determinar se uma palavra trata-se se um particípio ou adjetivo?

Por sua vez, Maria Helena Moura Neves, em sua Gramática de usos do Português (2011), menciona os particípios poucas vezes. A autora diz que "a locução verbal de voz passiva é formada com o verbo SER e o particípio do outro verbo", ressaltando o caráter verbal de particípios precedidos pelo verbo "ser", mas afirma também que é possível formar "uma voz passiva que indique estado usando-se o auxiliar ESTAR" (2011:65). Ela também afirma que adjetivos "terminados por sufixos que formam derivados de verbos, como -do/-to" são "prototipicamente predicativos" e "qualificadores" (2011:185). Ou seja, particípios podem formar voz passiva e assim ter um caráter verbal, como podem ser adjetivos — mas como distinguir os dois casos em contextos pouco claros?

A Gramática Pedagógica do Português Brasileiro, de Marcos Bagno (2011) é das raras obras que dedica algumas páginas à questão. Bagno trata particípios, infinitivos e gerúndios como uma classe nova, proposta por ele, os "verbinominais", e sugere alguns mecanismos para identificar quais particípios são /+verbo/ e quais são /+nome/". Porém, nos parece que o autor subestima a classe ao dizer ser "simples" fazer essa distinção (2011: 725).

Para Bagno, uma pista para a identificação de formais participiais / +verbo/ ou /+ nome/, quando antecedido de auxiliar, está nas propriedades flexionais: quando a forma participial é /+verbo/, não há flexão (*ele tinha comprado/ela tinha comprado*), e quando é /+nome/, há flexão (*ela estava acabada/ele estava acabado*). No entanto, os exemplos abaixo, provenientes do corpus NILC/São Carlos (Nunes et al., 1996), põem em xeque a

simplicidade da estratégia (ou da distinção) fornecida, uma vez que temos formas flexionadas e uma leitura que nos parece mais verbal:

(8): "Dada à própria origem do esporte, a palavra «basquetebol» é **resultada** do aportuguesamento da palavra inglesa «basketball»"

(9) "Na verdade, a origem da cultura grega então incipiente foi **resultada** de uma mescla de diversas heranças culturais anteriores"

(10): "O italiano Vico sustentou a tese de que a obra homérica foi **resultada** de vários poetas"

Bagno (2011) propõe ainda uma hipótese relacionando uso e o caráter /+verbo/ ou /+ nome/ de certas formais participiais. Para os verbos em que há mais de uma forma participial disponível em português, uma regular e outra (ou outras) irregular (aceitar – aceitado/aceito; acender – acendido/aceso; eleger – elegido/eleito; ganhar – ganhado/ganho; limpar – limpado/limpo etc), o uso teria feito com que as formas regulares mantivessem o caráter /+verbo/, deixando para as formas irregulares o caráter /+nome/, que, por sua vez, levariam a um emprego predominantemente adjetivo (Bagno, 2011:720).

Ao longo do processo de conversão das etiquetas do particípio que realizamos, porém, foi possível verificar que os dados não confirmam essa hipótese. A maioria dos particípios irregulares citados por Bagno aparecia no corpus do Mac-Morpho ou com ocorrências de ambas as leituras (aparecendo hora com leitura /+verbo/, ora com leitura /+nome/, dependendo do contexto) ou com ocorrências sempre verbais. Por exemplo:

(11) "A acusação atinge a Gráfica Gazeta de Alagoas, **acusada** de sonegar R$ 122 mil e de ter aceito notas fiscais frias"

(12) "A prestação de contas do partido foi **entregue** ontem à noite ao TRE (Tribunal Regional Eleitoral)"

(13) "E foi definitivamente **expulso** do Olimpo"

Além disso, Bagno apresenta alguns critérios sintáticos para mostrar se a leitura do *verbonominal* será adjetival ou verbal. Reproduzimos a seguir a parte referente aos particípios do quadro publicado nas páginas 723-724 da

Gramática, no qual são apresentadas, para cada frase, a análise sintática correspondente, com o objetivo de mostrar a aplicação das estratégias para a distinção entre as formas /+verbo/ ou /+ nome/. Uma leitura cuidadosa, no entanto, evidencia que a análise – não problematizada pelo autor – é exatamente o ponto crucial da questão, uma vez que é a partir dela que se estabelece o caráter /+verbo/ ou /+ nome/ dos chamados verbonominais.

PARTICÍPIO	+verbo	• Apaixonado, eu?	núcleo de sentença simples
		• Ana não tem viajado muito.	verbo principal de tempo composto
		• Passado o carnaval, voltamos para Brasília.	núcleo de sentença subordinada (reduzida)
		• Eu acho a Ana muito folgada.	núcleo de minissentença [▸600]
	+nome	• Achado não é roubado.	substantivo: sujeito / adjetivo: predicativo
		• Ana tem dez anos de casada.	substantivo: adjunto nominal preposicionado
		• Ana, descontraída, ia feliz pela rua.	adjetivo: adjunto nominal

Figura 3: Análise /+verbo/ ou /+ nome/ de verbonominais — particípios. Bagno (2011:723-724)

Não está claro, por exemplo, por que em "Eu acho a Ana muito **folgada**" estamos diante de uma forma /+verbo/ , mas em "Ana, **descontraída**, ia feliz pela rua" estamos diante de uma forma /+nome/. Os resultados apresentados no capítulo 5 desta dissertação corroboram a dificuldade de uma análise única, consensual, para certas formas participais.

No campos dos estudos linguísticos, a situação não é muito diferente da encontrada nas gramáticas: apesar da classificação dos particípios ser reconhecidamente um ponto problema[13], não há muitos estudos sobre o assunto. Margarida Basílio (2004) considera as formas –do adjetivos quando não há verbo auxiliar: "O sufixo -do se adiciona virtualmente a qualquer verbo para formação do Particípio Passado que, na forma variável, pode ser

[13] Veja-se, por exemplo, o seguinte trecho de Villalva e Silvestre (2014:164), grifo nosso: "Problemas suscitados pela classificação de palavras como o infinitivo, o gerúndio e o particípio (...) indicam que as classes de palavras tradicionalmente reconhecidas constituem matéria que pode e deve ainda vir a ser discutida".

utilizado quer na formação da voz passiva […], quer na adjetivação pura e simples" (Basílio, 2004:58). Basílio, diz ainda que a formação de adjetivos "correspondentes a particípios passados vai além da utilização da voz passiva" (2004:57), e utiliza o exemplo "Ela quer casa, comida e roupa **lavada**" para exemplificar o que seria um desses adjetivos sem vínculos com voz passiva — o que achamos bastante curioso, pois "lavada" nos parece ter uma clara procedência passiva, dependendo do contexto para que se possa afirmar se atua como adjetivo ou verbo.

Há alguns poucos trabalhos relativamente recentes abortando a questão, como Freitas et al. (2006), Oliveira & Freitas (2006) (sendo esses dois já no âmbito da linguística computacional) e Foltran & Crisóstimo (2012), mas nenhum deles foge muito às ideias presentes em Pimenta-Bueno (1986) — indubitavelmente o trabalho de referência na área dos estudos dos particípios — nem apresenta soluções novas para o problema da classificação desse grupo de palavras.

Pimenta-Bueno (1986) propõe que particípios passados sejam classificados em três grupos: adjetivos, verbos e particípios passivos. De acordo com ela, particípios funcionarão como verbos apenas quando precedidos pelos auxiliares "ter" ou "haver". Nos casos em que o particípio atua em estruturas com uma leitura passivo-eventiva, casos considerados pela autora como híbridos, ela sugere que não sejam considerados nem verbos nem adjetivos, mas uma terceira classe, que chama de particípios passivos. Em todos os outros casos, Pimenta-Bueno defende que os particípios sejam considerados adjetivos, e justifica sua posição com o argumento de que os particípios apresentam as seguintes propriedades, compartilhadas com adjetivos[14]:

1) Particípios podem ocorrer em posição predicativa (como, por exemplo, na frase "Hélio era **assustado** quando garoto");

2) Podem aparecer dentro do sujeito, tanto como núcleo como em outras posições (como em "Janelas **fechadas** fazem mal à saúde");

[14] Todos os exemplos apresentados com as propriedades foram retirados de Pimenta-Bueno (1986).

3) Podem ocorrer em expressões comparativas (como na frase "Márcia ficou tão **amolada** com a morte de D. Glorinha quanto o José");

4) Podem ocorrer em expressões superlativas (como em "Funaro é o mais **conhecido** dentre todos os Ministros da Nova República");

5) Apresentam formas superlativas absolutas sintéticas (como, por exemplo, em "João anda **agitadíssimo** e nervosíssimo e vive correndo de um lado para o outro");

6) Podem ser modificados pelos advérbios "bem", "muito" e "bastante" (como em "Este tópico é bastante **conhecido**");

7) Podem acontecer em coordenação com adjetivos, mas não com verbos (como na frase "Como estas crianças estão nervosas e **agitadas**!");

8) Concordam em gênero e número com o substantivo a que se referem (vide os exemplos anteriores).

Pimenta-Bueno ressalta que todas essas propriedades dos particípios se aplicam também aos adjetivos, mas não aos verbos, reforçando a semelhança entre as duas primeiras categorias. Ela afirma, porém, que os particípios podem apresentar também duas outras propriedades, que não são compartilhadas com adjetivos, mas são com alguns verbos. São estas:

9) Podem ocorrer imediatamente após um verbo e antes de um substantivo (como, por exemplo, na frase "Marta Rocha foi **coroada** 'Miss Brasil' na década de 50");

10) Podem ocorrer imediatamente após um verbo e antes de um adjetivo (como em "Leonardo foi **considerado** totalmente incapaz para o cargo").

Quando os particípios ocorrem em estruturas como 9 e 10, Pimenta-Bueno os considera casos híbridos, chamando-os de *particípios passivos* (PP).

As oito primeiras propriedades elencadas pela autora para demonstrar a semelhança entre particípios e adjetivos são convincentes, e talvez por isso sejam replicadas até hoje praticamente sem contestação (Foltran & Crisóstimo, 2005; Freitas et al., 2006). Porém, ao buscar por

particípios em grandes corpora, logo se observa que a questão é mais incerta do que o artigo de Pimenta-Bueno faz parecer e que essas regras na realidade se revelam insuficientes como critério de classificação para todos os casos de particípios, porque nem todos os particípios se comportam da forma descrita/esperada, como discutimos a seguir. Uma das dificuldades, parece, está no que poderiam ser passivas sem auxiliar[15]:

(14) "A produção de um vídeo, **feito** com objetividade e respeito, mostrou com duro realismo a situação de alguns centros educacionais que abrigam crianças, adolescentes e adultos portadores de deficiência"

(15) "Durante a festa, apareceu um cantor **vestido** de Elvis Presley"

(16) "Ainda que se discuta o valor, o «peso» literário dos já **citados** e dos outros escritores do programa e da antologia -- afinal, que representatividade terão Domingos Pellegrini, Marina Colasanti e tantos outros? , o fato é que à Alemanha interessa outra coisa"

(17) "Ela depois me mandou um cartão agradecendo, e disse que o que mais tinha chamado a a atenção dela fora a palavra «ornament», que não era muito **usada**"

Na frase (14), "feito" não se aplica a nenhum critério mencionado por Pimenta-Bueno – a não ser o critério 8. Não é um particípio precedido por verbo auxiliar, não está em uma construção verbal (propriedades 9-10) e nem em uma construção totalmente adjetival, conforme as propriedades 1-8 (não aceita, por exemplo, uma construção superlativa, como "o vídeo mais feito com objetividade e respeito", ou uma construção comparativa tipo "um vídeo tão feito com objetivo e respeito quanto aquele filme"). Apesar de não haver um agente explícito, há uma leitura passiva para "feito", de forma que o classificamos como um verbo. Pimenta-Bueno, porém, não explicita seu posicionamento em relação a esse tipo de estrutura.

Em relação à frase (15), esta também não parece se encaixar perfeitamente nos critérios mencionados. Não há um particípio precedido por verbo auxiliar, não se encaixa nas propriedades 9-10 e nem em uma

[15] Todas as frases listadas vieram do corpus Mac-Morpho.

construção totalmente adjetival, conforme as propriedades 1-8. Nesse contexto, "vestido" não se encaixa nas propriedades[16] 1, 3, 4, 5 e 6.

Os exemplos (14) e (15), apesar de não estarem precedidos por verbos auxiliares nem aparecerem em construções explicitamente passivas, não parecem apresentar um comportamento tão adjetival quanto palavras como "preocupado", por exemplo, o que nos obriga a repensar a generalidade de tais critérios. A possibilidade de intensificação parece estar mais vinculada à semântica do adjetivo– ser graduável ou não – do que à classe dos adjetivos.

Em (16), o particípio "citados" aparece em uma construção caracteristicamente mais verbal do que adjetival. Além do advérbio "já" carregar uma noção de temporalidade, a estrutura parece tratar-se da oração reduzida de "o peso literários dos *que já foram citados*". A critério de ilustração: ao buscar no corpus Mac-Morpho exemplos de "já" combinado com verbos (formas não participais), encontramos 829 resultados, enquanto as ocorrências de "já" combinado com adjetivos (formas não participais) são apenas 36; fizemos também a mesma busca no AC/DC e obtivemos 103.300 ocorrências para "já" + verbo e 11.633 para "já" + adjetivo (sendo grande parte destes resultados formas participiais), e na Floresta Sintá(c)tica, onde encontramos 7.709 ocorrências para a primeira busca e 313 (idem) para a segunda.

Já o exemplo (17) é um caso interessante, pois o particípio "usada" está inclusive numa construção de modificação adverbial ("muito usada"), mas está também participando de uma construção que pode ser interpretada como verbo auxiliar + verbo ("não *era* muito *usada*"). Além disso, a palavra "usado" apresenta dois sentidos levemente diferentes para a mesma forma, um como o particípio derivado da forma verbal "usar", e outro já

[16] Sobre as propriedades 3-5: uma breve pesquisa no corpus e até mesmo no Google mostra que "vestido" não é uma palavra frequentemente modificada por "muito" ou passível de formação superlativa absoluta, como em "apareceu um cantor muito vestido de Elvis Presley", ou "um cantor vestidíssimo de Elvis Presley", ainda que tais construções sejam sempre possíveis, como no caso de "gravidíssima".

cristalizado como o adjetivo “usado” (por exemplo, em “o carro não está sendo usado” é claramente distinto de “vendo roupas novas e usadas”).

Todos esses exemplos mostram que os critérios elencados não são suficientes para dar conta dos casos com que nos confrontamos.

Além disso, como esperamos ter ficado evidente ao longo desta seção, a literatura sobre o tema apresenta posicionamentos muito discrepantes, não havendo muita concordância entre os autores (por exemplo, enquanto Pimenta-Bueno tende a considerar grande parte dos particípios como adjetivos, Cunha e Cintra tendem a considerá-los como verbos).

3.2.2
Soluções empíricas e do PLN – o que fazem os corpora anotados

Nesta seção, apresentamos como três diferentes corpora anotados e revisados do português – Bosque, Mac-Morpho e Corpus UD[17] – lidam com as formas participiais.

Como informa Sampson (2001), corpora anotados não deixam de ser a materialização de uma gramática, e por isso consideramos relevante incluí-los aqui. Nesse contexto, é a documentação que explicita a gramática, mas é importante lembrar que nem sempre todos os fenômenos anotados estão explicitados na documentação. Nesse caso, é apenas a partir da observação dos exemplos que poderemos inferir a “filosofia gramatical” subjacente à anotação.

A documentação do corpus Bosque (Freitas e Afonso, 2008), não menciona explicitamente o caso das formas participais. No entanto, uma varredura pelas formas –do no corpus indica claramente uma posição sistemática: as formas –do serão sempre consideradas particípios (formas de verbo, portanto), exceto nos seguintes casos:

[17] Estamos chamando de Corpus UD o corpus disponibilizado pelo projeto UD e disponível em https://github.com/UniversalDependencies/UD_Portuguese-BR

a) Quando não há verbo correspondente (como nos casos de "indesejado" ou "insaturado", por exemplo, que não derivam dos verbos — inexistentes — "indesejar" e "insaturar", mas sim das formas participiais "desejado" e "saturado")

b) Quando as formas claramente se referem a um substantivo (por exemplo: "Sabe por que eu amo tanto você, querido_N?")

Isso significa que palavras que claramente nos parecem adjetivos são sistematicamente consideradas particípios, como

(19) "(...) onde Leopold Bloom surge como simbiose do Ulisses «polytropos» (muito **viajado** e de muitas manhas)(...)";

(20) "(...) peixes **grelhados**";

(21) "**Animado** com um teste de vestiário, o tcheco Zdenek Zeman(...)"

Já a documentação[18] relativa à anotação do corpus Mac-Morpho trata especificamente do particípio. Nela, a posição adotada parece ser, de certa forma, a mesma do Bosque:

> "Devido à dificuldade em resolver a ambigüidade que pode ocorrer entre uma forma terminada em -do (a) dos verbos, que pode exercer tanto a função de adjetivo quanto do particípio de um verbo, dependendo do papel que este desempenha na sentença, decidiu-se criar uma etiqueta única e específica para tais casos, ou seja, toda vez que houver a ocorrência de um particípio em uma sentença, este receberá esta etiqueta, independente de exercer uma ou outra função." (página 24)

Uma diferença, nos parece, é que o Bosque opta por manter a informação dos particípios como uma especificação dos verbos – em termos estritos, portanto, pode-se dizer que as formas participiais, no Bosque, são verbos. No Mac-Morpho, a decisão é por explicar, no âmbito das classes, a forma PCP.

[18] Disponível em http://nilc.icmc.usp.br/macmorpho/macmorpho-manual.pdf

Como ambos os corpora derivam da anotação automática do parser PALAVRAS, é possível que a posição tomada tenha sido influenciada pelas opções do PALAVRAS.

No entanto, embora parecidas, a anotação de ambos os corpora não é idêntica. O Bosque é mais consistente nas suas escolhas, o que leva a opções questionáveis como demonstramos acima. O Mac-Morpho, por outro lado, é um pouco mais flexível, e formais participiais que claramente consideraríamos adjetivos são, de fato, anotados como adjetivos (conforme evidenciado pela frase "Segundo Márcio Santos, Rocha está **apavorado**_ADJ com a hipótese de não mais jogar em esta Copa"). A desvantagem da opção é que nem sempre é fácil, como temos visto, decidir com relação às leituras mais verbais ou mais nominais. No quadro abaixo, mostramos as opções do Mac-Morpho e do Bosque para a mesma palavra, "querido". Como há sobreposição de conteúdo entre os materiais, as frases de fato se repetem.

Mac-Morpho	Bosque
Considere o caso de nosso vizinho próximo e muito **querido**_ADJ, o presidente…	Que o meu **querido**_ V-PCP amigo João Carlos Espada tenha acesso...
Resolvo caminhar sozinho por lugares que me são **queridos**_PCP e...	Modelo muito **querido**_V-PCP ao dr. Álvaro Cunhal
Põe o blazer que em Paris já é outono, querido_PCP	**Querido**_V-PCP camarada Erich Honecker
Claro, meu **querido**_N,..	Romário era **querido**_V-PCP, mas em certos bares ele não entrava
Sabe por que eu amo tanto você, **querido**_PCP?	Sabe por que eu amo tanto você, **querido**_N?

Quadro 1: comparação da anotação da palavra "querido" no corpus Mac-Morpho e no corpus Bosque.

Por fim, o que chamamos de Corpus UD-PT. Trata-se de um corpus disponibilizado pelo projeto UD, mas sobre o qual há pouquíssima informação disponível. Sabemos apenas que foi revisado, e que as frases,

coletadas da internet, estão aleatorizadas[19]. Quanto à documentação, é a mesma referente ao projeto UD. Com relação aos particípios, também não há qualquer problematização especial, e o projeto apresenta as seguintes informações:

> "Participle is a non-finite verb form that shares properties of verbs and adjectives. Its usage varies across languages. It may be used to form various periphrastic verb forms such as complex tenses and passives; it may be also used purely adjectively. Other features may help to distinguish past/present participles (English), active/passive participles (Czech), imperfect/perfect participles (Hindi) etc."[20]

> "Note that participles are word forms that may share properties and usage of adjectives and verbs. Depending on language and context, they may be classified as either VERB or ADJ."[21]

Embora de qualidade, algumas inconsistências são evidentes na análise dos particípios (como veremos mais adiante), mas em termos gerais podemos inferir que as seguintes decisões nortearam a anotação:

i) As chamadas "passivas sem auxiliar" são anotadas como V. Exemplo:

(22) "Em 2007, o STF aceitou denúncia contra os 40 suspeitos de envolvimento no suposto esquema **denunciado**_V em 2005 pelo então deputado federal Roberto Jefferson"

ii) Formas participiais modificando N, e sem agente da passiva (nem leitura passiva), são consideradas ADJ. Exemplos:

(23) "É o mais **atrevido**_ADJ escândalo de desvio de dinheiro público flagrado_VERB na história do Brasil";

---

[19] O corpus está disponível em https://github.com/UniversalDependencies/UD_Portuguese-BR

[20] http://universaldependencies.org/u/feat/VerbForm.html, acessado em 12/06/2016

[21] http://universaldependencies.org/u/pos/ADJ.html, acessado em 12/06/2016

(24) “Com Jadson muito **apagado**_ADJ e Denilson e Casemiro ineficientes na saída de bola”;

(25) “Está num local **privilegiado**_ADJ de acessibilidade a nível rodoviário”.

iii) Formas participiais precedidas dos auxiliares ficar/estar são consideradas ADJ. Exemplos:

(26) “Os pesquisadores ficaram 12 dias **acampados**_ADJ na esperança de capturar o animal […]”;

(27) “No entanto, como ressaltei, essas iniciativas não estavam **articuladas**_ADJ […]”

iv) Formas participiais precedidas do auxiliar ter/ser são consideradas V. Exemplo:

(28) “A carga não foi **prejudicada**_V”

Apresentamos, por fim, uma pertinente observação de Kilgarriff (2012), que não apenas reconhece a complexidade da questão e as consequências da pouca atenção dada às formas participiais especialmente no contexto da lexicografia computacional, mas também indica que a situação que relatamos não é exclusiva do português:

> “One recurring area of difficulty, in all the languages for which we have been involved in lexicography—two recent examples being Polish and Estonian—is participles/gerunds. In English, most -ed forms can be verb past tenses or past participles, or adjectival, and -ing forms can be verbal, adjective, or gerunds; comparable distinctions apply to most European languages. In theory, it may be possible to distinguish the form (verbal participle) from the function (verbal, adjectival, or nominal) but the theory still leaves the lexicographer with a judgement to make: should the -ing form get a noun entry, or should the -ed form get an adjective entry? POS-taggers are stuck with the same quandary: Where they encounter an -ing form, should they treat it as part of the verb lemma, as an adjective, or as a noun? The problem has two parts: some syntactic contexts unambiguously reveal the function (the painting is beautiful; he was painting the wall) but many do not (I like painting; the painting school). But this is only the first problem. The second problem is that some gerunds and participial adjectives are lexicalized, deserving their own entry in the dictionary, and others are not: thus we can have the manoeuvring is beautiful and there is no question that

> manoeuvring is functioning as a noun, but there is also no question that it is not lexicalized and does not need its own dictionary entry. The upshot is that many word sketches contain verb lemmas which would ideally not be there, because they are the result of lemmatization of adjectival participles and gerunds, which should have been treated as adjective and noun lemmas in their own right."
>
> (KILGARRIFF & KOSEM, 2012)

3.3
Palavras denotativas em foco

Outra etiqueta que rendeu desafios, porém de mais simples resolução, foi o alinhamento entre as "palavras denotativas", presente apenas no Mac-Morpho e ausente no tagset UD.

As palavras denotativas, ao contrário dos particípios — que ficam entre duas classes —- apresentam o problema de (aparentemente) não se encaixarem plenamente em classe alguma. Chamadas por Kury (1960) de "palavras de difícil designação", assemelham-se ora a advérbios, ora a conjunções, ora a preposições, sem se encaixarem bem em nenhuma dessas classes, e muitas vezes não se assemelham em nada a nenhuma outra classe (Pereira, 1995).

Ainda de acordo com Pereira (1995) entre as posições mais comumente atribuídas às palavras denotativas estão as de: (i) pertencentes a uma subclasse dos advérbios, (ii) não pertencentes a nenhuma classe e (iii) formadores de uma classe à parte (podendo esta carregar o nome de "palavras denotativas" ou não), sendo defendida a ideia de que tais palavras não podem ser atribuídas a outras classes por não possuírem as características necessárias e apresentarem atributos únicos e distintos do que pode ser encontrado nas outras classes gramaticais.

A discussão a respeito das palavras denotativas não é das mais populares, uma vez que engloba palavras consideradas de pouca relevância teórica, de forma que não há um grande número de estudos ou posicionamentos a respeito. Por outro lado, são palavras de grande frequência na língua, e portanto é relevante que recebam um tratamento

sistemático. Informações relativas a essas palavras são encontradas principalmente em gramáticas tradicionais, embora nem todas reconheçam a sua existência. De fato, retomando a gênese das classes de palavras apresentada no início deste capítulo, vemos que diferentemente dos particípios, não há vestígios, nas classes gregas, do que viria a ser as "palavras denotativas". Na NGB, a única menção a essa "categoria" é uma observação na seção dos advérbios: "certas palavras, por não se poderem enquadrar entre os advérbios terão classificação à parte. São palavras que denotam exclusão, inclusão, situação, designação retificação, afetividade, realce, etc."; porém, não é explicitado o motivo exato para essas palavras não poderem ser consideradas advérbios. Resumimos a seguir, em ordem cronológica, as principais colaborações à literatura referente a esse tema.

Oiticica (1940 apud Pereira, 1995) foi um dos gramáticos que mais se dedicou a desenvolver conteúdo relativo às palavras denotativas. Ele sugere que sejam consideradas uma classe à parte, defendendo que são "inclassificáveis" nas classes tradicionais. Ele acredita que os autores que as distribuem entre advérbios e preposições o fazem sem muito critério, criticando tal prática. Oiticica sugere ainda que a classe das palavras denotativas apresente dezessete subclassificações, como aditivas, afetivas, aproximativas, afirmativas, designativas, exclusivas, explicativas, inclusivas etc.

Já Câmara Jr. (1964 apud Pereira, 1995) não dedica tanta atenção ao assunto, mas chama esse tipo de palavra de "partícula de realce" ou "partícula expletiva", pois costumam ser utilizadas para produzir ênfase, e o autor defende que o objetivo dessas partículas é apenas a expressividade. No entanto, quando levamos em conta palavras ou expressões como "também" ou "somente", anotadas no Mac-Morpho como palavras denotativas, não está claro em que momento a expressividade se manifesta.

Ali (1970 apud Pereira, 1995) sugere que as palavras denotativas sejam incluídas na categoria "expressões de situação", criada por ele, que abarca expressões espontaneamente produzidas no decorrer da fala. Ele

critica a utilização do termo "expletiva" para se referir a esse tipo de palavra, e rassalta a importância das expressões de situação na comunicação.

Macambira (1987 apud Pereira, 1995), por sua vez, explicitamente considera as palavras denotativas uma subclasse dos advérbios, aceitando que advérbios modifiquem substantivos. Esse posicionamento é polêmico e considerado problemático por muitos autores, dentre eles Oiticica (1940) e Pereira (1995).

Bechara (1991) trata brevemente das palavras denotativas em uma pequena seção em um capítulo sobre advérbios.

Pereira (1995) alinha-se à ideia de Oiticica (1940) de colocar as palavras denotativas em uma classe própria de mesmo nome, mas discorda da necessidade de criar subdivisões, considerando que tamanha precisão classificatória confunde mais do que ajuda. Reforça, porém, a necessidade de se saber distingui-las semântica, sintática e morfologicamente das outras classes gramaticais, para que não haja sobreposições ou casos nebulosos.

Por fim, Rocha Lima (1999) e Cunha & Cintra (2001) apresentam posicionamentos parecidos, não considerando as palavras denotativas exatamente uma classe a parte, mas sim palavras/locuções indicativas de afirmação, explicação, negação, exclusão, avaliação etc.

3.4
Classes gramaticais no PLN: POS e tagsets

Os tipos mais comuns de anotação referem-se à anotação de lema, morfossintática (anotação das classes gramaticais de cada token/palavra), anotação sintática parcial e completa, e anotação semântica (por exemplo, entidades mencionadas, papeis semânticos dentre outros).

Fazer a anotação de POS, ou etiquetar um corpus com classes de palavras, é geralmente uma etapa inicial no processamento computacional de uma língua. Uma curiosidade interessante é que a sigla POS vem do

inglês “part of speech”, ou “parte do discurso”, justamente a terminologia utilizada pelos gregos antigos para elaborar classificações que vimos no início deste capítulo.

Por sua vez, a segmentação de um texto em unidades básicas (tokens) – tarefa em geral anterior à anotação de POS – não é trivial, como se poderia imaginar inicialmente, já que a identificação de unidades pode ser motivada por alguns fatores distintos, como semântica, morfologia ou grafia, nem sempre coincidentes. De fato, a discussão linguística subjacente à identificação do construto teórico “palavra” (Biderman, 2001) é exportada para o contexto do PLN.

A filosofia do Bosque, por exemplo, prioriza as unidades semânticas e morfossintáticas, de forma que nomes próprios e expressões multovocabulares (MWEs) são tokenizados como uma única palavra (por exemplo, na sentença “A situação tende a se agravar, uma_vez_que nenhuma de as partes parece mostrar disposição de recuar”, a expressão “uma vez que” foi considerada uma única unidade, quando poderia em outra anotação ser considerado duas). No Mac-Morpho, as expressões multivocabulares também são levadas em consideração, mas a forma de anotação é um pouco diferente: apesar das palavras não estarem formalmente concatenadas[22], têm seus elementos marcados com uma mesma etiqueta — por exemplo, a expressão “uma vez que” aparece como três tokens, mas todos estão com a mesma etiqueta — uma_KS vez_KS que_KS — para indicar que se trata de uma única unidade). Assim, o que parece é que, no corpus em si, a tokenização segue, por um lado, o critério

[22] É importante ressaltar, porém, que há uma inconsistência entre a documentação e a anotação do Mac-Morpho: no manual do Mac-Morpho, são mencionadas diversas vezes unidades polilexicais, que seriam aproximadamente equivalente a MWEs. De acordo com o manual, essas expressões deveriam estar unidas e com uma única etiqueta (por exemplo “uma=vez=que_KS”). Porém, no corpus não é isso que se encontra; não há palavras formalmente concatenadas formando uma unidade e com uma única etiqueta, mas sim palavras separadas etiquetadas com a mesma etiqueta, conforme foi descrito acima. Como não há nenhuma ocorrência desse tipo de concatenação no corpus disponibilizado, optamos por relatar nesta dissertação a filosofia que encontramos no corpus em si, não na retratada no manual.

da palavra ortográfica (espaços em branco são considerados delimitadores), mas as etiquetas atribuídas seguem as unidades semânticas.

Seguindo o mesmo exemplo, a expressão “uma vez que” aparece no corpus de português desenvolvido pelo UD como “uma_DET vez_NOUN que_CONJ”. Apesar de os três tokens formarem uma expressão, estão separados e cada um apresenta uma etiqueta diferente. Porém, as MWEs não são ignoradas: os corpora do UD apresentam diversas camadas de informação, e em uma dessas camadas indica-se quando os tokens fazem parte de uma expressão multivocabular. Assim, “uma”, “vez” e “que” podem aparecer etiquetados de forma independente, mas também está presente na anotação a informação de que esses elementos fazem parte da MWE “uma vez que”. Ou seja, como a anotação é feita em camadas, na camada POS o que vale é o critério ortográfico, mas há uma camada apenas para a indicação de expressões multivocabulares. O exemplo aparece da seguinte forma no corpus:

1	Apesar	_	ADV	ADV	_	2	mwe	_	_
2	disso	_	ADP	ADPPRON	_	5	adpmod	_	_
3	,	_	.	.	_	2	p	_	_
4	Maluf	_	NOUN	PNOUN	_	5	nsubj	_	_
5	acredita	_	VERB	VERB	_	0	ROOT	_	_
6	que	_	CONJ	CONJ	_	9	mark	_	_
7	a	_	DET	DET	_	8	det	_	_
8	sociedade	_	NOUN	NOUN	_	9	nsubj	_	_
9	esta	_	VERB	VERB	_	5	ccomp	_	_
10	mais	_	ADV	ADV	_	11	advmod	_	_
11	atenta	_	ADJ	ADJ	_	9	acomp	_	_
12	nesta	_	ADP	ADP	_	9	adpmod	_	_
13	eleição	_	NOUN	NOUN	_	12	adpobj	_	_
14	,	_	.	.	_	19	p	_	_
15	uma	_	DET	DET	_	16	mwe	_	_
16	vez	_	NOUN	NOUN	_	17	mwe	_	_
17	que	_	CONJ	CONJ	_	19	mark	_	_
18	já	_	ADV	ADV	_	19	advmod	_	_
19	passou	_	VERB	VERB	_	9	advcl	_	_
20	por	_	ADP	ADP	_	19	adpmod	_	_
21	este	_	DET	DET	_	22	det	_	_
22	tipo	_	NOUN	NOUN	_	20	adpobj	_	_
23	de	_	ADP	ADP	_	22	adpmod	_	_
24	situação	_	NOUN	NOUN	_	23	adpobj	_	_
25	no	_	ADP	ADP	_	19	adpmod	_	_
26	pleito	_	NOUN	NOUN	_	25	adpobj	_	_
27	de	_	ADP	ADP	_	26	adpmod	_	_
28	2008	_	NUM	NUM	_	27	adpobj	_	_
29	.	_	.	.	_	5	p	_	_

Figura 4: exemplo de expressão multivocabular no corpus de português UD.

Tais considerações são relevantes porque, a uma dada segmentação, corresponde uma certa classificação, de forma que a segmentação e a

anotação de POS são interdependentes. Assim, para de fato converter uma anotação de forma a seguir determinado padrão, pode ser necessário realizar alterações também na tokenização.

Criar um tagset — ou seja, decidir quais classes gramaticais serão utilizadas para etiquetar um determinado corpus — pode a princípio parecer uma tarefa simples, já que as gramáticas tradicionais já apresentam listas com as classes gramaticais existentes e alguma descrição sobre elas. Porém, a tarefa está longe de ser trivial: as categorias apresentadas pelas gramáticas, na maioria das vezes, não são o suficiente para lidar com todas as produções de língua que podem ser encontradas em um corpus, as descrições oferecidas para cada classe muitas vezes não são o suficiente para resolver casos ambíguos, e não há consenso, entre as gramáticas, sobre quais seriam as classes utilizadas. Além disso, ao lidar com corpus, há a constante necessidade de lidar com fenômenos considerados periféricos pela tradição linguística, que não são abordados nas gramáticas, ou são apenas brevemente mencionados (como é o caso dos particípios e das palavras denotativas).

Tagsets podem ser bem diferentes entre si, a começar pelo tamanho: quanto maior o tagset, maior o nível de refinamento (ou granularidade) das classes e, consequentemente, das análises necessárias para a atribuição dessas classes. Manning & Schutze (1991) ressaltam que às vezes tagsets apresentam distinções de granularidade em algumas áreas e não em outras, dependendo do que é considerado mais relevante pela equipe desenvolvedora do tagset, e que algumas classes de palavras podem ser abordadas de formas completamente distintas em tagsets diferentes. Etiquetas mais granulares podem refletir distinções importantes capazes de fornecer informações úteis sobre os outros tokens daquele contexto, mas tornam a tarefa de classificação muito mais complexa, tanto para humanos quanto para máquinas. Quanto mais classes, mais potencial para a discordância.

Quando há a ambição de se alinhar determinados tagsets, não há muitas dificuldades em alinhar as classes mais convencionais sólidas, como substantivos e preposições, mas quanto mais granular for a descrição, mais abstrato será o nível de anotação linguística, e menos consensual será a classificação (Manning & Schutze, 1991).

Para Leech & Wilson (1996), o ideal é que se elabore um "esquema" de anotação (disponível ao público, para que os usuários do corpus possam compreender satisfatoriamente a anotação feita) tão preciso que, se dois anotadores diferentes forem utilizar o mesmo esquema para anotar um corpus, devem chegar a exatamente o mesmo resultado. Porém, na prática, há limites muito tênues e por vezes incertos entre as classes de palavras, o que pode gerar o que os autores chamam de "ambiguidades".

Para os autores, as ambiguidades tendem a surgir quando anotadores humanos têm dificuldade em decidir sobre qual etiqueta aplicar uma palavra. Eles explicam que isso pode ocorrer quando o esquema de anotação não apresenta critérios claros para desambiguação, quando dois ou mais anotadores humanos têm opiniões diferentes — ou perspectivas teóricas diferentes — sobre os dados, ou quando as próprias categorias apresentam delimitações pouco claras (o que acontece com certa frequência).

Achamos relevante destacar que, frequentemente, tais situações de discordância são fruto do caráter interpretativo da tarefa de anotação, nem sempre explícito – qualquer que seja a sua natureza. A anotação é sempre a interpretação de algo, que permite a inclusão desse "algo" em uma certa classe. No caso específico das classes de palavras, é razoável que haja discordância quanto à interpretação de um dado elemento, em um certo contexto – as discussões relativas às formas participiais, apresentadas na seção anterior, ilustram esse ponto. Por isso, nos parece razoável que, em certos contextos, mais de uma interpretação (mais de uma leitura) seja possível – o que não quer dizer que seja desejável no contexto de anotação de um corpus.

4
Metodologia

4.1
O corpus Mac-Morpho

O corpus Mac-Morpho (Aluísio et al., 2003), é um corpus de 1.1 milhão de palavras, desenvolvido no projeto Lacio-Web e revisado manualmente. O corpus é composto de textos em Português brasileiro, retirados de edições do jornal Folha de São Paulo de 1994 e anotados com POS. Os textos utilizados são de gênero predominantemente jornalísticos extraídos de dez seções do mencionado jornal. Para este trabalho, foi utilizada a primeira versão do corpus, disponibilizada em http:// nilc.icmc.usp.br/macmorpho, em três partes.

O lançamento do Mac-Morpho foi em 2003 e desde então duas revisões foram feitas, eliminando ruídos e fazendo alterações no tagset, gerando assim os corpus v. 2 (Fonseca & Rosa, 2013) e v. 3 (Fonseca et al., 2015). Na primeira revisão (v. 2) foram eliminadas frases repetidas e com palavras faltando e mudou-se a tokenização em relação a contrações (que na versão original eram desfeitas e anotadas separadamente), deixando-as como um único token e consequentemente adicionando uma nova etiqueta ao tagset (PREP+ART). Por exemplo, na v. 1, temos "em_PREP o_ART", enquanto na v. 2 esse tipo de ocorrência foi modificada para "no_PREP +ART". Na segunda revisão (v. 3), mais sentenças problemáticas (repetidas/ com palavras faltando) foram retiradas e houve novas mudanças no tagset, com a remoção de etiquetas que dependiam de interpretações acima do nível morfossintático. As etiquetas removidas foram "verbo auxiliar", "pronome conectivo relativo" e "advérbio conectivo relativo", que foram reanotadas com as etiquetas mais gerais "verbo", "pronome conectivo" e "advérbio conectivo", respectivamente.

Para a presente empreitada, optamos por utilizar o Mac-Morpho v. 1, por ser a versão mais compatível com a tokenização e o tagset do projeto

UD (que desconcatena contrações e apresenta a etiqueta “verbo auxiliar” em seu tagset). Após finalizado o progresso de conversão, porém, comparamos o corpus que geramos com o Mac-Morpho v. 3, a fim de eliminar as sentenças que foram removidas durante as duas revisões e obter um produto final igualmente limpo.

O tagset utilizado na v. 1 do Mac-Morpho é este:

Tagset do corpus Mac-Morpho
adjetivo ADJ
artigo ART
advérbio ADV
advérbio conectivo subordinativo ADV-KS
advérbio conectivo subordinativo relativo ADV-KS-REL
conjunção coordenativa KC
conjunção subordinativa KS
interjeição IN
moeda CUR
nome N
nome próprio NPROP
numeral NUM
palavra denotativa PDEN
particípio PCP
preposição PREP
pronome pessoal PROPESS
pronome adjetivo PROADJ
pronome conectivo subordinativo PRO-KS
pronome conectivo subordinativo relativo PRO- KS-REL
pronome nominal PROSUB
pronome pessoal PROPESS
verbo V

verbo auxiliar VAUX
. , ; : ? ! “ ‘ ( ) { } < > / [sem etiqueta]

Quadro 2: tagset do corpus Mac-Morpho.

4.2
Tagset “Universal” e projeto “Dependências Universais”

O projeto das Dependências Universais (UD) tem como objetivo desenvolver uma anotação multilingue de corpora consistente para diversas línguas. Isto é, criar tagsets — para POS e para dependências sintáticas — independentes de língua. De acordo com as informações fornecidas na página de apresentação do UD[23], o projeto busca gerar um conjunto universal de categorias e orientações que facilitem a anotação consistente de estruturas similares entre as línguas, mas também permitindo extensões específicas de determinada língua quando necessário. O esquema de anotação é baseado em uma evolução das dependências de Stanford (de Marneffe et al., 2006, 2008, 2014), que afirmam também serem universais, no tagset de POS universal do Google (Petrov et al., 2012) e na interlingua Interset para tagsets morfosintáticos (Zeman, 2008). Consideramos interessante apontar que na própria página de apresentação há a seguinte observação, em consonância com o trabalho desenvolvido nesta dissertação:

> “As a result of this work, universal POS categories have substantive definitions and are not necessarily just equivalence classes of categories in underlying language-particular treebanks. Hence, work to convert to UD POS tags often requires context-sensitive rules, or some hand correction.”

No momento[24], o projeto conta com 58 corpora anotados para as línguas Alemão, Antigo Eslavo Eclesiástico Aramaico, Árabe, Basco,

[23] http://universaldependencies.org/introduction.html, acessado em 15/07/2016

[24] Em julho de 2015.

Búlgaro, Catalão, Cazaque, Chinês, Cóptico, Coreano, Croata, Tcheco (três corpora), Dinamarquês, Esloveno (dois corpora), Espanhol (dois corpora), Espanhol Estoniano, Finlandês (dois corpora), Francês, Galego, Grego, Grego Antigo (dois corpora), Gótico, Hebraico, Hindi, Holandês (dois corpora), Húngaro, Indonésio, Inglês (três corpora), Irlandês, Italiano, Japonês, Latim (dois corpora), Letão, Norueguês, Persa, Polonês, Português (dois corpora), Romeno, Russo (dois corpora), Sueco (dois corpora), Tâmil, Turco, Ucraniano e Vietnamita. Esses corpora são corpora já existentes que foram convertidos automática ou manualmente para os padrões e para o tagset do projeto UD, ou corpora desenvolvidos no âmbito do Google. Os corpora já convertidos de Português são o Bosque, da Floresta Sintá(c)tica (Afonso et al., 2012, tomando por base a versão usada no CONLL em 2006), convertido automaticamente e um corpus desenvolvido pelo próprio Google, e disponível em https://github.com/UniversalDependencies/UD_Portuguese-BR, sobre o qual quase não há documentação. O Bosque, porém, é um corpus pequeno, com menos de 200 mil palavras, e os fatos de o corpus do Google não possuir documentação, ter as frases aleatorizadas e apresentar anotação inconsistente dificultam que este seja considerado referência. O corpus Mac-Morpho, além de também ser revisado, é amplamente utilizado no treino de sistemas em português brasileiro, e por isso decidimos investir na sua conversão. Além de mais material disponível em um contexto que parece promissor para o PLN de diferentes línguas, podemos também investigar o impacto na aprendizagem de diferentes tagsets.

As etiquetas de POS utilizadas nessa empreitada são as do tagset universal (Petrov et al., 2011), um tagset desenvolvido por Petrov, Das & McDonald (2011), a partir da ideia de que seria possível estabelecer categorias gramaticais comuns e gerais às línguas. Num momento inicial, o tagset foi apresentado com 12 etiquetas, mas esse número foi depois expandido para 14 — ou 17, contando com as etiquetas de pontuação,

símbolos e fragmentos desconhecidos. As etiquetas do tagset universal são as seguintes:

Tagset do projeto UD
ADJ adjetivo
ADP adposição
ADV advérbio
AUX verbo auxiliar
CONJ conjunção coordenativa
DET determinante
INTJ interjeição
NOUN substantivo
NUM numeral
ADV advérbio
PRON pronome
PROPN nome próprio
PUNCT pontuação
SCONJ conjunção subordinativa
SYM símbolo
VERB verbo
X outros

Quadro 3: tagset do projeto UD.

4.3
Criação de regras e demais aspectos técnicos da conversão

Para realizar a conversão, nós criamos um conjunto de regras gerais (cerca de 60) e muitas regras específicas ou "exceções" (cerca de 3 mil)[25], isto é, regras estabelecidas para dar conta de casos específicos. Para a aplicação das regras, foi desenvolvido um script utilizando a linguagem de programação Lisp[26]. Quando havia etiquetas diretamente equivalentes nos

[25] O script de regras está disponível em https://github.com/own-pt/macmorpho-UD sob o nome "ud-remove-pcp".

[26] Agradecemos imensamente a Alexandre Rademaker e a Fabrício Chalub por todo o auxílio no desenvolvimento do programa.

tagsets, a tarefa era simples: escrevíamos um comando para que todas as ocorrências de uma etiqueta (por exemplo, "_PREP") fossem substituídas pela equivalente (no caso, "_ADP"). Porém, quando não havia equivalência direta, o procedimento era mais complicado e significativamente mais trabalhoso.

Para converter os PCPs, por exemplo, primeiro pegamos uma amostra de 200 casos em que essa etiqueta aparecia e lemos todas as frases em busca de padrões que pudessem virar regras gerais de conversão.

Elaboramos então algumas regras e as ordenamos de forma que a ordem não afetasse o funcionamento da regra. Por exemplo: colocamos como nossa primeira regra "*_VAUX sido_PCP *_PCP" vira "*_VAUX sido_PCP *_V", o que quer dizer que em construções do tipo "algum verbo auxiliar + particípio 'sido' + algum particípio", o último particípio sempre será transformado em verbo. A segunda regra é "sido_PCP *_V" vira "sido_VAUX *_V", ou seja, nos casos em que ocorrer " particípio 'sido' + verbo", o particípio 'sido' será convertido em verbo auxiliar. Pode-se perceber que a segunda regra é dependente da primeira, de forma que as duas precisam ser aplicadas nessa ordem para que a conversão funcione adequadamente.

Procuramos também colocar as regras mais genéricas antes das com menos poder de generalização, ou seja, as primeiras regras não apresentam exceções, depois há algumas regras com poucas exceções e então regras com exceções cada vez mais numerosas, até o caso de "regras" que se aplicam uma única vez.

Cada vez que escrevíamos uma regra geral, testávamos utilizá-la no corpus e líamos absolutamente todas as ocorrências (que às vezes chegavam a 7 mil) em busca de exceções. Sempre que encontrávamos uma, transformávamos aquela ocorrência em uma regra específica, que era então redigida antes da regra geral. Por exemplo: nossa regra 5 é uma regra geral, que consiste na transformação da sequência "*_VAUX *_ADV *_PCP" em "*_VAUX *_ADV *_V", o que quer dizer que quando há uma construção

com um verbo auxiliar, um advérbio e um particípio, o particípio será um verbo. Porém, na frase "o_ART time_N foi_VAUX quase_ADV perfeito_PCP", "perfeito" claramente não é um verbo. Assim, criamos uma regra específica antes da regra 5 (geral) para transformar essa ocorrência específica em "o_ART time_N foi_VAUX quase_ADV perfeito_ADJ", de forma que esse PCP será transformado em ADJ antes que regra geral seja aplicada, escapando dela.

Durante nossa busca por exceções, sempre que encontrávamos alguma inconsistência ou algum erro de anotação em qualquer parte do corpus, criávamos uma regra para corrigi-lo. Dessa forma, nossa conversão final foi feita em uma versão revista do corpus[27]. Disponibilizamos, inclusive, essa versão do Mac-Morpho com essa revisão e seu próprio tagset. Mais detalhes sobre a revisão do material serão abordados no próximo capítulo.

Quando encontrávamos uma palavra que era sempre convertida para uma mesma etiqueta, independentemente do contexto, também criávamos uma regra que consista na conversão da palavra para essa etiqueta (como por exemplo a palavra "acostumado", que sempre aparecia como adjetivo, e "feito", que sempre aparecia como verbo). Todas as regras desse tipo estão dentre as regras iniciais do script, para não permitir a interferência de outras regras.

Nós também optamos por identificar nos scrips, utilizando etiquetas que não se manifestam no corpus, quando uma regra era referente a) a uma correção, c) a uma expressão multivocabular e d) a um erro de digitação. Para isso, utilizamos as tags "cr", "mwe" e "typo", respectivamente. A figura 5 ilustra algumas regras:

```
(-> "faz_V sentido_PCP" "faz_V sentido_N" :cr)
(-> "Assim_KC como_KC" "Assim_ADV como_PREP"  :mwe :misc "MWE=Assim_como|MWEPOS=CONJ")
(-> "no_KC entando_KC" "no_PREP entando_ADV" :typo  :mwe :misc "MWE=no_entanto|MWEPOS=CONJ")
```

Figura 5: exemplo de regras utilizadas com etiquetas "cr", "mwe" e "typo".

[27]O script de regras está disponível em https://github.com/own-pt/macmorpho-UD sob o nome "ud-keep-pcp".

Depois de finalizada a conversão, lemos amostras aleatórias do corpus para conferir se tudo estava de acordo com o planejado, fazendo alterações nas regras quando necessário. Mas, devido ao tamanho do corpus, é possível que alguns erros ou inconsistências possam ter passado por nossa revisão.

Além disso, geramos também um corpus com o tagset UD, porém mantendo a etiqueta "particípio". Para a criação desse corpus, mantivemos todas as regras de conversão que não envolvessem a conversão de particípios[28]. Esse ponto será detalhado no próximo capítulo.

4.4
Correção de erros no Mac-Morpho

Durante o nosso trabalho, sempre que encontrávamos alguma anotação errada, criávamos uma regra no script de programação para corrigi-la. Alguns exemplos das correções feitas (no total, foram feitas por volta de 400 correções[29]):

(73) "Combina pistola a laser, c**oletor_N de_PREP dados_PCP** e transmissor de radiofrequência", foi corrigido para "coletor_N de_PREP dados_N";

(74) "**Piloto_N fez_N** primeiro teste em a F-1 em Imola" foi corrigido para "Piloto_N fez_V";

(75) "[…] por isso delineou um outro teste, em que camundongos previamente estressados com exposição a **forte_N ruído_PCP** eram colocados em amplas gaiolas abertas" foi corrigido para "forte_ADJ ruído_N";

---

[28] O script de regras está disponível em https://github.com/own-pt/macmorpho-UD sob o nome "mm-revisto".

[29] As correções estão disponíveis em https://github.com/own-pt/macmorpho-UD/blob/master/mm-revisto.lisp, marcadas com a etiqueta ":cr".

(76) “O fotógrafo escolhe então quais **serão_V enviadas_N**, conecta o Photolynx a uma linha telefônica e começa a transmissão” foi corrigido para “serão_V enviadas_PCP”.

Também optamos pela padronização de anotações que nos pareceram inconsistentes. Essas correções foram feitas não por serem “erros” propriamente ditos, mas por não estarem de acordo com o padrão do Mac-Morpho. Por exemplo, está no manual do Mac-Morpho a informação de que “quando o particípio de um verbo aparece como núcleo em um sintagma nominal, este receberá então a etiqueta de nome e não de particípio”. Esse posicionamento fica evidente em exemplos como:

(77) “Enquanto os bois criados em regime de pasto perdem peso por falta de alimentação, os **confinados_N** adquirem até 1,5 quilo a o dia.”

(78) “Mas os retardatários incorrerão em multa de 100% e os **omissos_N** serão lançados de ofício com os dados disponíveis e com multa de 300% .”

Assim, os PCPs que ocupavam posição de substantivo foram por nós transformados em Ns, como o do exemplo seguinte:

(79) “os **desempregados_PCP** continuarão sendo cerca de 12% de a população economicamente ativa”.

Consideramos importante ressaltar que não corrigimos erros de digitação (como por exemplo, a ocorrência de “envaidas” ao invés de “enviadas”), por considerá-los parte do corpus. Porém, marcamos todos os que encontramos com a tag “:typo”, para facilitar a correção caso seja interessante fazê-lo eventualmente.

4.5
Metodologia de avaliação das decisões linguísticas

Durante este trabalho, por vezes foi necessário tomar decisões de anotação sobre as quais não há consenso na literatura — particularmente no caso dos particípios. Para avaliar e validar nossas decisões linguísticas, considerando as divergências de análise, desenvolvemos um teste usando a

ferramenta Rêve (Santos et al. 2015). O Rêve foi desenvolvido no contexto da Gramateca (Santos, 2014), com a proposta de "possibilitar ao linguista um ambiente de teste de hipóteses on-line, gratuito e público" (Santos et al., 2015), viabilizando também o estudo de diferentes interpretações ou discordâncias de anotação. A ferramenta possibilita a criação de questionários ou experimentos online, utilizando o material disponível na Linguateca, e apresentando total transparência quanto às perguntas, às respostas e aos resultados. Dessa forma, avaliamos nossas decisões linguísticas por meio das concordâncias (ou discordâncias) de opinião entre especialistas. A seção 5.2.4. detalhará esse processo de validação. O argumento subjacente é o da busca do consenso, considerando a inexistência de uma resposta (ou análise) considerada correta na literatura linguística.

4.6
Metodologia de avaliação do impacto dos tagsets

Para avaliar o impacto dos diferentes tagsets apresentados nesta pesquisa, utilizamos um mesmo sistema para anotar automaticamente os datasets que geramos. Com essa estratégia, sabemos que as diferenças nos resultados obtidos se devem unicamente às variações nos tagsets.

Além da comparação dos valores de acurácia do sistema com os diferentes datasets, procedemos a uma análise dos erros considerando os dados gerados em cada matriz de confusão. Esta etapa teve um duplo objetivo: (a) investigar com mais detalhe o impacto das variações dos tagsets e (b) servir como fonte para a revisão da própria anotação inicial (revisão do golden), levando, portanto, a um aprimoramento do recurso, uma vez que a análise dos erros evidencia também quando os erros não são fruto de análises erradas, mas de erros na análise humana original.

As informações sobre o sistema utilizado, os resultados de acurácia com cada tagset e as análises de erros a partir dos dados obtidos são apresentados em mais detalhes no capítulo 6.

5
Construção de datasets

5.1
Conversão dos tagsets

Neste capítulo, serão detalhadas as etapas que levaram à conversão do Mac-Morpho para o tagset UD e à construção de dois datasets: o Mac-Morpho com o tagset UD e o Mac-Morpho com o tagset UD acrescido da etiqueta de particípio.

5.1.1
Alinhamento

Após cuidadosa análise dos tagsets e busca de exemplos de cada classe nos corpora, chegamos ao seguinte alinhamento entre as etiquetas do Mac-Morpho e do Google:

Mac-Morpho	Google
adjetivo ADJ	ADJ adjetivo
advérbio ADV	ADV advérbio
preposição PREP	ADP adposição
verbo auxiliar VAUX	AUX verbo auxiliar
conjunção coordenativa KC	CONJ conjunção coordenativa
artigo ART pronome adjetivo PROADJ	DET determinante
interjeição IN	INTJ interjeição
nome N	NOUN substantivo
numeral NUM	NUM numeral
palavra denotativa PDEN	ADV advérbio ou expressão multivocabular com tags intependentes para cada elemento
pronome pessoal PROPESS pronome nominal PROSUB	PRON pronome
nome próprio NPROP	PROPN nome próprio

. , ; : ? ! “ ‘ ( ) { } < > / [sem etiqueta]	PUNCT pontuação
conjunção subordinativa KS pronome conectivo subordinativo PRO-KS pronome conectivo subordinativo relativo PRO- KS-REL advérbio conectivo subordinativo ADV-KS advérbio conectivo subordinativo relativo ADV-KS-REL	SCONJ conjunção subordinativa
moeda CUR	SYM símbolo
verbo V	VERB verbo
particípio PCP	VERB verbo, ADJ adjetivo, NOUN substantivo ou ADV advérbio
-	X outros
-	PRT partícula

Quadro 4: alinhamento entre as etiquetas do Mac-Morpho e do projeto UD.

Como é possível observar, ainda que em boa parte dos casos o alinhamento seja simples, alguns casos se mostraram especialmente trabalhosos. Algumas etiquetas têm equivalentes diretos, como o “V” (verbo) do Mac-Morpho e o “VERB” (verbo) do UD. Outras etiquetas do Mac-Morpho podem ser aglomeradas sob uma única etiqueta do UD, como é o caso de “PROPESS” (pronome nominal) e “PROSUB” (pronome pessoal), que viram ambos “PRON” (pronome) na conversão. Porém, há casos mais complexos, em que não é possível fazer uma conversão direta e torna-se necessário desmembrar uma única etiqueta do Mac-Morpho em duas ou mais etiquetas do UD. É o caso das etiquetas “PCP” (particípios) e “PDEN” (palavras denotativas).

Em relação às palavras denotativas, nas ocorrências em que há apenas uma palavra marcada como PDEN, esta palavra é convertida para ADV (advérbio) — a não ser nos casos marcados como erros de anotação[30],

[30] A medida que o alinhamento foi sendo feito, quando percebíamos que havia erro na anotação original, aproveitamos para corrigir, cf. seção 4.4

que foram convertidos para algumas categorias distintas. Porém, quando há duas ou mais palavras marcadas como PDEN[31], formando uma expressão multivocabular, é necessário aplicar uma etiqueta diferente para cada elemento, para seguir o padrão UD. Assim, torna-se impossível fazer a conversão por meio de uma única regra: é necessário buscar todas as ocorrências de palavras denotativas e convertê-las caso a caso.

Algo semelhante ocorre com os particípios: a etiqueta "PCP" não apresenta equivalente no tagset UD, de forma que as palavras assim anotadas devem ser distribuídas entre verbos, adjetivos, substantivos e, vez ou outra, até mesmo advérbios. É possível criar algumas regras gerais de conversão (por exemplo, no Mac-Morpho, particípios precedidos por verbo auxiliar sempre serão verbos), mas tais regras dão conta de apenas alguns tipos de particípio; a maioria dos casos não se mostrou passível de padronização a ponto de possibilitar a criação de regras gerais de conversão, tornando-se necessária novamente a análise caso a caso.

5.2
Desafios linguísticos da conversão

Ao reanotar um corpus com um tagset diferente do original, não basta somente alinhar as etiquetas do tagset original com as do novo. Ou seja, os alinhamentos apresentados na tabela 2 não são o resultado da conversão, mas uma das etapas do trabalho de conversão. A outra etapa consiste em verificar o alinhamento entre as concepções gramaticais subjacentes às etiquetas e à própria anotação. Assim, é também necessário levar em consideração a filosofia de anotação específica de cada corpus. Esta tarefa é desafiadora, pois nem sempre as decisões de anotação estão explicitadas – isto é, documentadas, como vimos no capítulo 3. Nestes casos, torna-se necessário buscar exemplos variados no corpus das

---

[31] Conforme explicamos anteriormente, no corpus Mac-Morpho, as expressões multivocabulares não aparecem concatenadas, mas têm seus elementos marcados com uma mesma etiqueta — a expressão "por exemplo" aparece como "por_PDEN exemplo_PDEN".

ocorrências em questão, a fim de depreender a filosofia por trás de uma anotação. Em resumo: porque a classificação gramatical não é uma tarefa desvinculada de teoria – e de interpretação – e porque são conhecidas as críticas e limitações da classificação gramatical, por um lado, e a ausência de uma proposta alternativa consensual, por outro, a conversão de um tagset não é apenas uma tarefa mecânica de conversão de etiquetas. Uma mesma etiqueta pode ser usada com finalidades diferentes, e por isso a relevância, na tarefa de conversão, da manutenção da filosofia de anotação: o fato de dois corpora serem anotados com o mesmo tagset não garante que estejam alinhados. A seção a seguir detalha os casos em que um alinhamento baseado apenas no nome da etiqueta leva a corpora com anotações distintas. Vale lembrar que uma consequência desse desalinhamento é a impossibilidade de usá-los como recursos complementares, por exemplo, ampliando o material de treino disponível.

5.2.1
Filosofias distintas de anotação

Conforme já foi dito, muitas vezes, dois tagsets apresentam uma mesma classe, mas isso não quer dizer que essas classes iguais tenham exatamente o mesmo comportamento gramatical e sejam usadas da mesma maneira em ambos os corpora. No caso específico da anotação de classes de palavras, ilustramos, no capítulo 3, o percurso acidentado desta classificação, desde a Grécia Antiga até chegar ao que temos hoje. Apresentamos abaixo os casos em que o alinhamento baseado apenas no "nome" da classe levaria a um corpus inconsistente do ponto de vista da anotação UD.

Números — NUM

O tagset do Mac-Morpho possui uma classe para números, cuja etiqueta é "NUM". O tagset UD tem exatamente a mesma etiqueta, com o mesmo significado. Porém, enquanto nos corpora anotados do projeto UD a

orientação do manual é de que números cardinais sejam marcados com a etiqueta "NUM", no corpus do Mac-Morpho apenas uma parcela está marcada como tal, devido à seguinte instrução:

> "Quando um numeral (ainda que composto apenas por números) funciona, em uma sentença, como núcleo do sintagma nominal (SN), muitas vezes regido por preposição, receberá a etiqueta de Nome (e não de Numeral)" (manual do Mac-Morpho[32])

O manual cita ainda os seguintes exemplos:

(1) "O crime aconteceu em **1978**_N."

(2) "Em **1990**_N, Carla terminou a faculdade."

(3) "O Brasil ganhou de **2**_N a **0**_N."

(4) "Entrem **um**_N de cada vez."

(5) "Entre **14**_N e **18**_N de abril, haverá festa no clube."

(6) "Em 14_N de abril Paula completará **40**_NUM anos."

(7) "Era **1o**_N de abril."

(8) "No ano de **1997**_N, Lucia formou-se em medicina."

Assim sendo, para converter o Mac-Morpho de acordo com o UD, foi necessário reanotar como "NUM" todas as ocorrências de numerais que no primeiro estão marcadas com a etiqueta "N".

Pronomes — PRON

O caso dos pronomes é um tanto delicado. A definição da etiqueta "pronome" (PRON) do projeto UD é a seguinte:

> "Pronouns are words that substitute for nouns or noun phrases, whose meaning is recoverable from the linguistic or extralinguistic context.

[32] http://nilc.icmc.usp.br/macmorpho/macmorpho-manual.pdf, acessado em 28/01/2016

> Pronouns under this definition function like nouns. Note that some languages traditionally extend the term pronoun to words that substitute for adjectives. Such words are not tagged PRON under our universal scheme. They are tagged as determiners in order to annotate the same thing the same way across languages.
> For instance, [en] this is either pronoun (I saw this yesterday.) or determiner (I saw this car yesterday.) Its Czech translation, [cs] tohle, is traditionally called pronoun in Czech grammar, regardless of context (the notion of determiners does not exist in Czech grammar). To make the annotation parallel across languages, it should be now tagged PRON in Tohle jsem viděl včera. and DET in Tohle auto jsem viděl včera."[33]

O alinhamento com base no nome da classe pode levar a um descompasso com o que se pretende na anotação. Por exemplo, temos no Mac-Morpho 5 etiquetas de pronomes: pronome conectivo subordinativo (PRO-KS), pronome conectivo subordinativo relativo (PRO- KS-REL), pronome pessoal (PROPESS), pronome nominal (PROSUB) e pronome adjetivo (PROADJ). Uma conversão automática e/ou incauta poderia alinhá-las todas à etiqueta "pronome" (PRON) do UD. Esse alinhamento, porém, não é recomendável, pois as etiquetas PRO-KS e PRO- KS-REL na verdade se encaixam melhor em "conjunção subordinativa" (SCONJ) e o alinhamento mais indicado para PROADJ seria "determinante" (DET). Apenas as etiquetas PROPESS e PROSUB de fato se encaixam em PRON.

Expressões multivocabulares — MWE

Outro exemplo que explicita muito bem diferenças filosóficas de anotação é o tratamento dado a expressões linguísticas, formadas por duas ou mais palavras (expressões multivocabulares, ou MWE),

Expressões multivocabulares se, por um lado, são cruciais para sistemas de PLN, por outro, são reconhecidamente difíceis de serem tratadas

[33] http://universaldependencies.org/docs/u/pos/PRON.html, acessado em 25/06/2016

(Ramish, 2012). De fato, a dificuldade na identificação de MWEs tem a ver com a dificuldade na delimitação do que seja palavra (Biderman, 2011). Considerando a complexidade do tema, não iremos aqui discutir os diferentes conceitos e critérios utilizados na identificação de uma MWE, mas apenas as diferentes maneiras de formalizá-las, nos projetos de anotação mencionados (Mac-Morpho e UD), ponto relevante no alinhamento entre os corpora.

Como um dos objetivos da nossa empreitada é justamente converter o Mac-Morpho para o padrão UD, utilizar a mesma solução do UD é a opção indicada. Assim, convertemos o Mac-Morpho para o formato CONLL, o que nos permite ter colunas adicionais para acrescentar informações relevantes. Dessa forma, ao lidar com expressões multivocabulares, optamos por etiquetar cada palavra individual e separadamente ("ou" marcado como conjunção e o "seja" marcado como verbo, por exemplo) e assinalar na coluna MWE que "ou seja" é uma expressão multivocabular equivalente a advérbio — exatamente como é feito no projeto UD.

Utilizamos essa estratégia para as expressões multivocabulares marcadas no Mac-Morpho como palavras denotativas (ex: "ou_PDEN seja_PDEN", que se tornou "ou_CONJ seja_V"), para aquelas marcadas como advérbios (ex: "ainda_ADV por_ADV cima_ADV", que se tornou "ainda_ADV por_PREP cima_ADV") e para as marcadas como conjunções (ex: "a_KC não_KC ser_KC", que se tornou "a_CONJ não_ADV ser_V"). Um desafio na identificação de MWEs no Mac-Morpho é saber quando duas etiquetas idênticas e consecutivas são a indicação de uma MWE ou correspondem apenas a palavras consecutivas de mesma classe. Ou seja, em certos casos é fácil identificar a MWE, já que "por" em "por exemplo" é uma preposição e nunca uma palavra denotativa, mas em casos como "logo após", por exemplo, a detecção não é tão simples. Apenas esses três tipos de concatenação de expressão multivocabular (PDEN, ADV e KC) aparecem na nossa conversão marcadas como "mwe".

Compostos com hífen

Compostos com hífen e palavras derivadas por prefixação com hífens, como "semi-elaborados_N", "ex-governador_N", "pró-feminina_ADJ", "vice-presidente_N", "pós-fixados_ADJ", dentre outros, também ilustram diferenças relacionadas a filosofias distintas de anotação. Palavras como "centro-sul_NPROP", "puro-sangue_ADJ", "infra-estrutura_N", "matéria-prima_N", "primo-irmão_N", "tiro-curto_N", "norte-americano_ADJ", dentre outros, estão assim anotados no Mac-Morpho: como uma única palavra, não havendo segmentação e colocando-se apenas uma etiqueta no conjunto.

Já em relação ao projeto UD, o posicionamento a respeito de formação com hífen não é explicitado. Esta é a única informação fornecida:

> "The universal dependency annotation is based on a lexicalist view of syntax, which means that dependency relations hold between words. Hence, morphological features are encoded as properties of words and there is no attempt at segmenting words into morphemes. However, it is important to note that the basic units of annotation are syntactic words (not phonological or orthographic words), which means that we systematically want to split off clitics, as in Spanish dámelo = da me lo, and undo contractions, as in French au = à le."[34]

Sem orientações explícitas sobre o que fazer em relação a palavras com hífen, recorremos aos corpora. Analisando exemplos do corpus UD de inglês, que consideramos o padrão (disponível em https://github.com/UniversalDependencies/UD_English), podemos encontrar ocorrências de, por exemplo:

[34] http://universaldependencies.org/u/overview/tokenization.html, acessado em 25/06/2016

(29) “Indo-Sri Lanka”
Indo_X
-_PUNCT
Sri_PROPN
Lanka_PROPN

(30) “self-rule”
self_NOUN
-_PUNCT
rule_NOUN

(31) “ill-advised”
ill_ADV
-_PUNCT
advised VERB

(32) “so-called”
so_ADV
-_PUNCT
called_VERB;

(33) “full-time”
full_ADJ
-_PUNCT
time_NOUN”;

(34) “wide-ranging”
wide_ADV
-_PUNCT
ranging_VERB

Ou seja, palavras compostas com hífen parecem ser separadas e anotadas de forma independente. Porém, também podemos encontrar ocorrências como:

(35) “mid-nineties_NOUN”;

(36) “cross-examination_NOUN”;

(37) “non-social_ADJ”;

(38) “pro-India_ADJ”;

(39) “anti-army_ADJ”;

(40) “neo-conservatives_NOUN”;

(41) “multi-millionnaires_NOUN”.

Essas ocorrências (dentre muitas outras), dão a entender que palavras formadas por derivação prefixal com hífen são mantidas e recebem apenas uma etiqueta.

Porém, esse padrão não é encontrado em todos os corpora. Por exemplo, no corpus UD-PT, a anotação não parece consistente com os outros corpora do UD, e o padrão de tokenização é sempre separar palavras com hífen, como vemos em:

(42) “Centro-Oeste”
Centro_PNOUN
-_.
Oeste_PNOUN

(43) “produtor-executivo”
produtor_NOUN
-_.
executivo_ADJ

(44) “vice-liderança”
vice_PRT
-_.
liderança_NOUN

(45) “primeiro-ministro”
primeiro_ADJ
-_.
ministro_NOUN”;

(46) “tornou-se”
tornou_VERB
-_.
se_PRT

(47) “tornou-se”
tornou_VERB
-_.
se_PRON

(48) “ex-amante”
ex_PRT
-_.
amante_NOUN

(49) “ar-condicionado”
ar_NOUN
-_.
condicionado_ADJ

(50) “co-autoria”
co_PRT
-_.
autoria_NOUN”.

Consideramos importante destacar que às vezes ocorrências idênticas recebem etiquetas distintas nesse corpus, como se pode verificar em “tornou_VERB -_. se_PRT” e “tornou_VERB -_. se_PRON”, o que indica que a anotação desse corpus não foi devidamente padronizada e que ele não é a melhor opção para ser utilizada como referência. Além disso, apesar da etiqueta “PRT” (partícula) aparecer em alguns desses contextos, ela não aparece na versão UD do Bosque e a documentação do UD relativa a essa etiqueta parece sugerir uma classe sem equivalentes para o português, o que nos leva a crer que os casos acima tratam-se de erros de anotação.

Ainda tomando por base a versão UD do Bosque, disponibilizada pelo projeto UD como padrão para o português[35], em geral, se mantém as opções do Bosque original, e palavras com hifen ficam juntas, como evidenciado em "Auto-Conhecimento_PROPN", "sexta-feira_NOUN", "mano-a-mano_NOUN", "dia-a-dia_NOUN", "preto-e-branco_NOUN", "vice-versa_ADV", "vice-presidente_NOUN", "pró-natalista_ADJ" e até mesmo "ex-vice-primeiro-ministro_NOUN", dentre muitos outros.

Devido a essa disparidade de exemplos encontrada nos corpora e à falta de tempo e de mão-de-obra para refazer a tokenização e anotação manualmente e caso a caso - a princípio, de acordo com o padrão do corpus do inglês (que, devemos lembrar, não está nem explicitado, foi deduzido por nós através dos exemplos) - optamos por manter o formato original do Mac-Morpho — nesse aspecto, idêntico ao do Bosque, já reconhecido pelo projeto UD.

5.2.2
Palavras denotativas

Ao reclassificar as palavras denotativas presentes no Mac-Morpho, levamos em consideração as definições de advérbio de Rosa (2002) e Ilari et al (1991), um pouco mais abrangentes do que as tradicionais. Rosa (2002) apresenta a classe dos advérbios como sendo uma classe não muito homogênea; são modificadores, mas não modificam substantivos, e sim verbos, adjetivos, outros advérbios, sintagmas e até mesmo a própria sentença. Ilari et al. (1991) expandem o conceito de advérbio, possibilitando que abarquem inclusive elementos de organização discursiva (como "então", "inclusive", "agora", "aí" etc). Baseamo-nos nessa definição para alinhar as palavras marcadas como "palavras denotativas" no corpus Mac-Morpho ao tagset UD, de forma que, ao nos depararmos com palavras que se encaixassem nessa caracterização, nós as convertíamos para advérbios. Buscamos também ocorrências de algumas dessas palavras no Bosque e no

[35] Disponível em https://github.com/UniversalDependencies/UD_Portuguese

corpus de inglês do projeto UD, e constatamos que elas aparecem marcadas como advérbios, corroborando nossa decisão. As únicas palavras originalmente etiquetadas como "PDEN" que não convertemos para advérbios foram aquelas que participavam de expressões multivocabulares. Conforme foi explicado no capítulo 4, no caso de expressões multivocabulares, optamos por seguir o padrão UD e etiquetar cada palavra independentemente, anotando em outra coluna a informação de se tratar de uma expressão equivalente a um advérbio.

5.2.3
Particípios

Em um corpus como o Mac-Morpho, que é composto por textos retirados da Folha de São Paulo, pode-se encontrar tudo aquilo presente em um jornal: algumas crônicas, entrevistas e até mesmo ocasionais contos, mas a maior parte do corpus é ainda assim formado por notícias. Textos jornalísticos, devido à necessidade de concisão e precisão, apresentam abundância de particípios, mas em condições distintas das mencionadas por Pimenta-Bueno, 1986, (explicadas em maiores detalhes no capítulo 3), como fica evidenciado nos seguintes exemplos (todos tirados do Mac-Morpho[36] — negrito nosso):

(51) "É a segunda medida de liberação econômica **anunciada** em menos de uma semana."

(52) "A família de o deputado João Batista (PSB), **assassinado** em dezembro de 88, teme uma fuga de o pistoleiro Péricles Ribeiro Moreira, 35, **acusado** do crime."

(53) "Com 5.774 unidades **comercializadas** em esse primeiro trimestre, a previsão é de vender 28 mil em 94."

[36] Todos os exemplos deste capítulo foram retirados do Mac-Morpho, a não ser que seja explicitado diferente.

(54) “O avião **roubado** no aeroclube dos Amarais, em Campinas, em o final de o mês passado, foi encontrado na última quarta-feira na Bolívia em poder de um homem identificado apenas por Juan Carlos.”

(55) “Relatório de Banco Mundial, **publicado** em 91 e assinado por Avishay Braverman e Monika Huppi, mostra que sempre o subsídio é uma boa opção para incentivar o aumento de plantio.”

Ao analisar os particípios em negrito com algum cuidado, percebe-se que grande parte das propriedades apresentadas por Pimenta-Bueno (1986), senão todas, não são aplicáveis nesses casos. Os usos mencionados por ela como verbais (precedidos por verbo auxiliar) ou participiais passivos (seguidos pela preposição “por”) também não se encaixam nesses exemplos.

Como, então, deve ser feita a classificação desses particípios? Devem ser considerados verbos ou adjetivos, apesar de não se aplicarem perfeitamente às propriedades propostas por Pimenta-Bueno (1986)?

Nesse momento da pesquisa, ficou claro que seria necessário tomar um posicionamento quanto à classificação dos particípios, estabelecendo critérios, que nos permitissem seguir com as classificações.

Depois de analisar literalmente milhares de exemplos[37] no Mac-Morpho e no UD, e muito os discutir, chegamos aos seguintes critérios, levando em consideração as anotações do UD (quando disponíveis):

i) Formas participiais precedidas pelos auxiliares ter/ser/haver são consideradas verbos. Exemplos:

(56) “Os dois carros **são vendidos**_PCP com ágio, em o mercado paralelo”

ii) Formas participiais precedidas pelos auxiliares ficar/estar são consideradas adjetivos.

(57) “O viaduto **ficou** completamente **interditado**_PCP até_PREP a as 9h10.”.

iii) Particípios em construções passivas com agente explícito seriam sempre considerados verbos. Exemplos:

---

[37] Foram alterados exatamente 23.093 casos, para os quais criamos mais de 3 mil regras/ exceções.

(58) “Segundo pesquisa **realizada**_PCP **por** o Datafolha em o último dia 25 de julho, Serra tem 30% de as intenções de voto, contra 27% de Erundina.”

iv) As chamadas “passivas sem auxiliar” são anotadas como verbo.

(59) “O caderno especial sobre os 10 anos de a derrota de a emenda que restabelecia eleições diretas, **publicado**_PCP em o domingo passado, conseguiu opiniões unânimes de os leitores”.

v) Particípios presentes em combinações convencionais são considerados adjetivos (como em, “semana passada”, “países desenvolvidos”, “revendedora autorizada”, “revistas especializadas”, dentre muitos outros[38]).

vi) Formas participiais modificando N, e sem agente da passiva (nem leitura passiva), são consideradas adjetivos.

(60) “Além de devoto, Ricupero é um **disciplinado**_PCP estudioso_**N** de religião.”

Em casos que não se aplicassem à nenhuma condição citada acima ou que fossem de difícil classificação (por vezes, pode ser complicado distinguir se uma sentença é iv ou vi, por exemplo, pois decidir se uma palavra tem valor passivo ou não pode depender da interpretação que se tem da frase), recorremos também às seguintes regras:

vii) Particípios que satisfizessem (a maioria d)as propriedades propostas por Pimenta-Bueno (1986) para identificar adjetivos seriam considerados adjetivos.

viii) Particípios que não satisfizessem (a maioria d)essas propriedades propostas por Pimenta-Bueno (1986) seriam considerados verbos.

No entanto, concluímos que os critérios que criamos para a aplicação das regras, apesar de bastante eficazes na maioria dos casos, não podem ser seguidos cegamente.

---

[38] A lista completa de combinações está contida no script de regras de conversão do particípio, disponível em https://github.com/own-pt/macmorpho-UD/blob/master/ud-remove-pcp.lisp.

Como vimos, há casos em que particípios apresentam comportamento similar tanto ao de verbos quanto de adjetivos, tornando a decisão classificatória complexa: por exemplo, em "a palavra era muito **usada** naquela época", a palavra "usada" se encaixa em nossa regra i), mas também satifaz as propriedades características de adjetivos de Pimenta-Bueno. Ao realizar um questionário com professores/pesquisadores da área de gramática (cf. seção deste capítulo 5.2.4.), a maioria esmagadora classificou a palavra como um verbo — e nós mesmas somos dessa opinião. Defendemos então, que em casos com esse (que não são poucos), o contexto e o valor semântico são de extrema relevância, representando sempre o maior peso na decisão da classificação de uma palavra.

Outro ponto digno de ser ressaltado é a questão da polissemia das formas participais, que muitas vezes é deixada de lado e é crucial para se compreender não só o motivo de alguns casos parecerem claros e se encaixarem tão bem em alguma categoria, como também a razão da inviabilidade de realizar uma conversão por meio da aplicação de alguma(s) regras(s) geral(is) de particípios para outras classes. Algumas formas participiais, ao longo do tempo, acabam se lexicalizando e ganhando um sentido específico. A palavra "importado", por exemplo, pode ser usada com o sentido de "estrangeiro" (sentido esse inclusive previsto em dicionário), não estando mais necessariamente associada ao ato de importar.

Nos exemplos abaixo temos o mesmo particípio sendo utilizado de três formas distintas (considerados pela anotação dos corpora Mac-Morpho e Português UD como adjetivo, como verbo, como verbo novamente e substantivo).

(61) "No entanto, o mesmo presidente da Abeiva vangloria-se da criação de 25 mil empregos diretos nas 620 empresas consessionárias de carros **importados**." (ADJ)

(62) "[…] um aumento de 30 pontos percentuais no Imposto sobre Produtos Industrializados (IPI) para carros **importados** ao Brasil de fora do

Mercosul" (exemplo do corpus de Português do Google UD criado a partir do Google Universal Dependency Treebanks 2.0) (V)

(63) "[…] o besouro foi **importado** da Alemanha para cá entre 1950 e 1959" (exemplo do corpus de Português do Google UD criado a partir do Google Universal Dependency Treebanks 2.0) (V)

(64) "[…] a opinião do presidente, ao diminuir a alíquota do IPI dos **importados**, […]" (N)

Como as regras propostas por Pimenta-Bueno não dão conta de todos os casos participiais que encontramos no corpus, a documentação do UD não problematiza especificamente particípios em português e as análises linguísticas divergem (como apresentamos no capítulo 3), optamos por elaborar nossas próprias soluções, fruto de nossa interpretação, em muitos casos. Algumas vezes, porém, fomos obrigadas a categorizar mesmo sem muita convicção, devido sobretudo à necessidade de transformar os particípios em V ou ADJ, por conta da tarefa de anotação.

Como maneira de validar nossas opções, fizemos um questionário sobre a classificação das formas participiais, que descrevemos na próxima seção.

5.2.4
Validação das decisões linguísticas relativas ao particípio

Como existe na literatura uma tendência a abordar os particípios majoritariamente como adjetivos (Pimenta Bueno, 1986; Basílio, 2009), e nós muitas vezes nos deparávamos com casos de particípios que considerávamos claramente verbais, decidimos consultar outros profissionais e pesquisadores da área de gramática, validar os critérios propostos e para verificar, nos casos difíceis — que foram analisados individualmente — como diferentes especialistas se posicionariam.

Além disso, conforme apontado por Santos et al. (2015), sempre há muito debate nos projetos associados à anotação humana quanto à

concordância entre anotadores. A concordância é vista como fator de validação das decisões de anotação e parece haver uma ideia implícita de que concordância é um sinônimo de qualidade, mas essa associação pode-se revelar equivocada, já que, quando anotadores erram em uma mesma direção, esses erros acabam virando concordâncias, sem que isso seja equivalente a uma anotação de qualidade. Os autores ressaltam ainda que discordância entre anotadores não precisa ser vista como indicativo de baixa qualidade, pois há áreas em que a interpretação e o contexto são muitíssimo relevantes, e fenômenos que exigem tomadas de decisão nada simples — como é o caso dos particípios, cujo caráter híbrido dificulta a escolha de uma classificação, podendo gerar altos índices de discordância entre anotadores.

Assim, criamos um questionário utilizando a ferramenta Rêve (Santos et al., 2015). O Rêve foi desenvolvido no contexto da Gramateca[39](Santos, 2014), com a proposta de "possibilitar ao linguista um ambiente de teste de hipóteses on-line, gratuito e público" (Santos et al., 2015), viabilizando também o estudo de diferentes interpretações ou discordâncias de anotação. A ferramenta possibilita a criação de questionários ou experimentos online, utilizando o material disponível na Linguateca, e apresentando total transparência quanto às perguntas, às respostas e aos resultados.

Cada participante deveria ler uma série de frases contendo uma palavra em negrito, sempre um particípio. Para cada frase, o participante deveria indicar se a palavra em negrito era, naquele contexto, um verbo ("V"), um adjetivo ("ADJ"), ou alguma outra classe ("Outra"). Caso o participante estivesse em dúvida ou não soubesse ao certo como classificá-la, poderia também selecionar a opção "Não sei". Era possível marcar mais de uma opção, criando combinações como "ADJ/V" ou "V/Não sei", por exemplo. Também havia, para cada frase, uma caixa de comentários que o participante poderia usar para explicitar os critérios que motivaram sua

[39] http://www.linguateca.pt/Gramateca/

escolha ou indicar dúvidas, caso desejasse. A figura abaixo exemplifica o funcionamento do questionário.

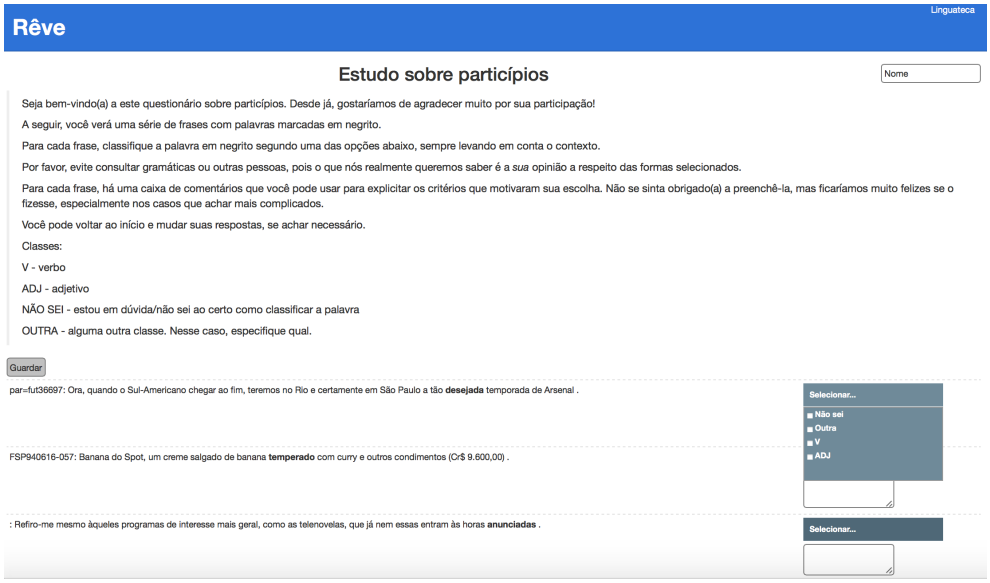

Figura 6: página do questionário no Rêve.

O questionário continha um total de 48 frases, distribuídas da seguinte maneira: 12 frases com particípios que consideramos mais claramente verbais, 12 com particípios que consideramos mais claramente adjetivais e 24 casos que consideramos difíceis, com particípios híbridos, que permitem uma leitura tanto adjetival quanto verbal, podendo depender dos critérios utilizados na análise. A ordem de apresentação das frases foi aleatorizada. Abaixo ilustramos o que consideramos "frases com particípios mais claramente verbais" e "frases com particípios mais claramente adjetivais", e frases que consideramos difíceis.

Frases com particípios que consideramos claramente verbais incluíam:

i) particípios precedidos por verbos auxiliares, como:

(65) "Edmundo negou e foi **chamado** de mascarado."

ii) particípios em construções passivas com agente explícito, como:

(66) "Outro problema **levantado** pelo pesquisador está com os produtores que pagam para colhedores autonômos."

iii) particípios que sem dúvida representavam orações reduzidas, como:

(67) "A também queniana Hellen Kimaiyo, 25, venceu a prova feminina, **disputada** mais cedo".

Frases com particípios que consideramos adjetivais consistiam em:

i) frases com particípios em expressões convencionais, como:

(68) "Nefasta ideia, que deveria estar totalmente erradicada, à luz de estudos recentes, publicados em prestigiosas revistas **especializadas**."

ii) particípios em posição pré-nominal ou pós-nominal sem complemento, cuja única função era caracterizar um substantivo, como:

(69) "As novas tecnologias de material de fundição e a discussão sobre técnicas de gestão para sobreviver em um mercado **globalizado** são os destaques do 7º Congresso de Fundição e da 6ª Feira de Fundidos, Insumos e Equipamentos, abertos ontem em São Paulo";

(70) "A família Kennedy pediu a William Manchester -- um **conhecido** jornalista, autor e historiador -- que escrevesse o que viria a ser a versão autorizada do assassinato de JFK".

iii) particípios coordenados com adjetivos, como:

(71) "A vida era **fechada**, escura, estratificada".

iv) particípios desempenhando função de predicativo, como:

(72) "O corte do atacante Bettega, principal astro da equipe, também deixou o país **desconfiado**".

Participaram do questionário um total de 10 pessoas, todos professores universitários e/ou pesquisadores da área de Letras com experiência na área de gramática do Português. Decidimos convocar participantes com esse perfil para investigar se haveria consenso nas respostas para os casos que consideramos difíceis, ou se os profissionais também ficariam hesitantes e divididos, como nós ficamos. Nossa hipótese era de que, nesses casos, haveria mais divergência do que convergências nas respostas, reforçando o caráter híbrido das formas participiais.

As respostas dadas pelos participantes foram na direção que esperávamos: os particípios que consideramos originalmente como verbais e adjetivais de fato foram alvo de muito mais concordância do que aqueles que consideramos casos difíceis.

É interessante ressaltar que as respostas para os particípios verbais (gráfico 1) foram as mais homogêneas. As respostas para os particípios adjetivais (gráfico 2) também foram relativamente homogêneas, mas já houve maior divergência do que para os verbais, havendo inclusive uma frase com 50% de respostas desviantes da análise esperada:

(73) "Ficariam resguardadas desta efetiva desvalorização da moeda **indexada** dos ricos apenas as cadernetas de poupança".

Imaginamos que essa divergência tenha ocorrido porque "moeda indexada" pode ser vista como uma expressão específica da economia, talvez não conhecida por todos os participantes.

Já para os casos difíceis (gráfico 3), houve grande variação de respostas em praticamente todas as frases (só houve concordância total em duas frases). Achamos muito interessante essa variação de respostas para cada frase, pois todos os participantes tinham grande conhecimento e domínio de gramática. Uma primeira interpretação dessa variação diria que, se particípios fossem abordados satisfatoriamente na gramática tradicional e houvesse uma orientação clara em relação à classe gramatical a qual pertencem, os participantes certamente conheceriam tais orientações e se comportariam de forma similar no questionário, o que não ocorreu. Preferimos outra análise: estamos de fato diante de uma classe híbrida. Além disso, enfatizamos que os particípios não foram, em sua maioria esmagadora, considerados adjetivos (comportamento que seria previsto baseando-se nas propostas de Pimenta-Bueno, 1986) e que as análises dos participantes em relação à classificação dos particípios na maioria das vezes alinhava-se à nossa, fornecendo maior segurança para avançar com nossa classificação.

Calculamos a concordância da seguinte forma: como eram 10 participantes, cada resposta de um participante era computada como 10%. Como era possível marcar mais de uma resposta, no caso da combinação de duas respostas, 5% era computado para cada uma (por exemplo, se em determinada frase 9 participantes respondessem "V" e um participante respondesse "V/ADJ", os resultados desse particípio ficariam 95% V e 5% ADJ). Apenas 4 participantes usaram classificações múltiplas.

Apresentamos então os dados detalhados que obtivemos com o questionário:

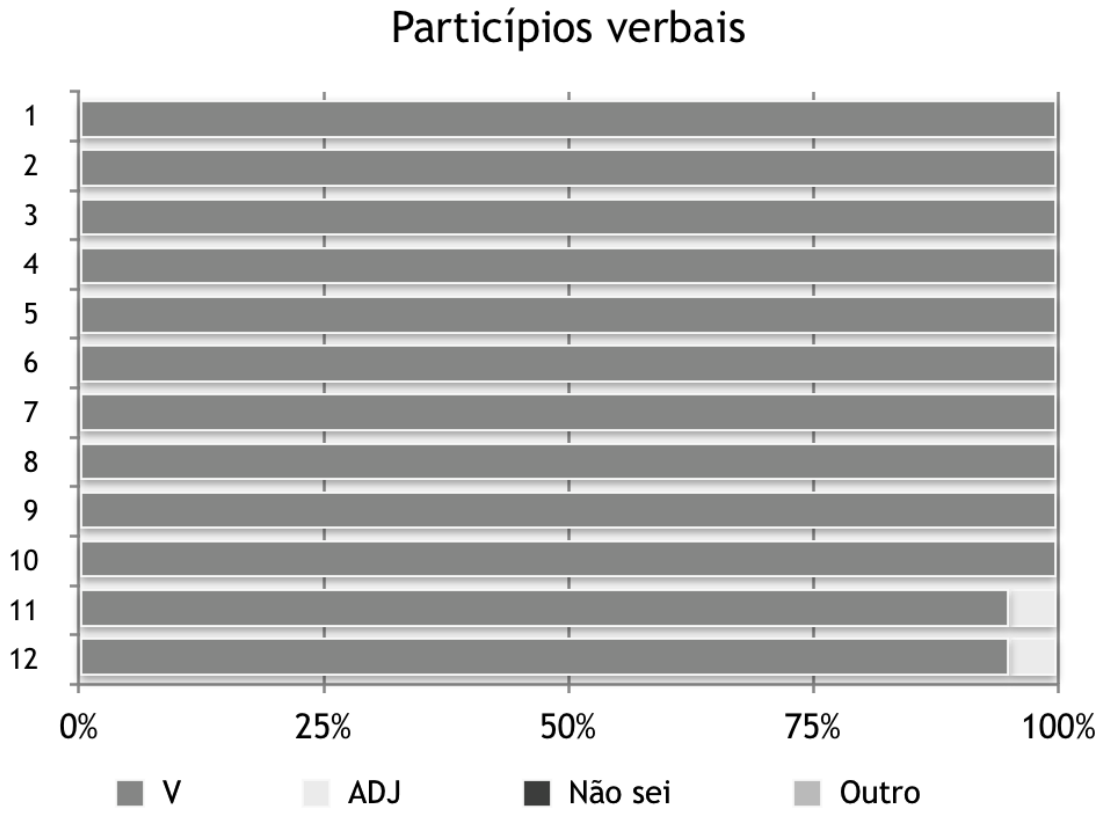

Gráfico 1 - Particípios verbais considerados fáceis

Frase 1 (100% V): "Edmundo negou e foi **chamado** de mascarado."

Frase 2 (100% V): "Suas apresentações foram feitas no último final de semana em um megarecinto recentemente **construído** com capacidade para 20 mil pessoas e 80 camarotes."

Frase 3 (100% V): "Esta é a conclusão da Polícia Federal do Amazonas depois de prender uma menina de 16 anos com 12 kg de cocaína

e apreender 4 kg da droga **deixados** no aeroporto de Tabatinga (AM) por um menino."

Frase 4 (100% V): "A também queniana Hellen Kimaiyo, 25, venceu a prova feminina, **disputada** mais cedo."

Frase 5 (100% V): "O robô, **desenvolvido** em quatro anos, usa chips de computador como código genético e peças previamente montadas como células."

Frase 6 (100% V): "A hemoglobina produzida pela Baxter e baptizada DCLHb («Diaspirin Crosslinked Hemoglobin») é o resultado de uma transformação química daquela molécula natural **obtida** através da utilização de um derivado da aspirina."

Frase 7 (100% V): "Escrito em menos de três meses, de janeiro a abril de 1774, publicado no mesmo ano e **traduzido** para o francês já em 1775, o livro foi mal recebido pela crítica, fazendo, entretanto, sucesso imediato junto ao público (uma série de suicídios alastra-se pela Europa após sua publicação)."

Frase 8 (100% V): "O livro foi **editado** pela multinacional McCann Erickson -- Portugal"

Frase 9 (100% V): "A produção de um vídeo, **feito** com objetividade e respeito, mostrou com duro realismo a situação de alguns centros educacionais que abrigam crianças, adolescentes e adultos portadores de deficiência."

Frase 10 (100% V): "Das buscas que a GNR de Albufeira fez à vivenda onde residia, a «Casa Anas», em Santa Eulália, e no escritório foram apreendidos diversos documentos relacionados com os veículos que se presume terem sido **furtados** no Algarve ou no estrageiro, mas todos com matrículas inglesas, falsas."

Frase 11 (95% V, 5% ADJ — 9 participantes marcaram V, 1 participante marcou V/ADJ): "Outro problema **levantado** pelo pesquisador está com os produtores que pagam para colhedores autonômos."

Frase 12 (95% V, 5% ADJ — 9 participantes marcaram V, 1 participante marcou V/ADJ): "Num incidente noturno **registrado** na época passada, a polícia foi encontrá-lo sentado no seu camião com uma pistola descarregada."

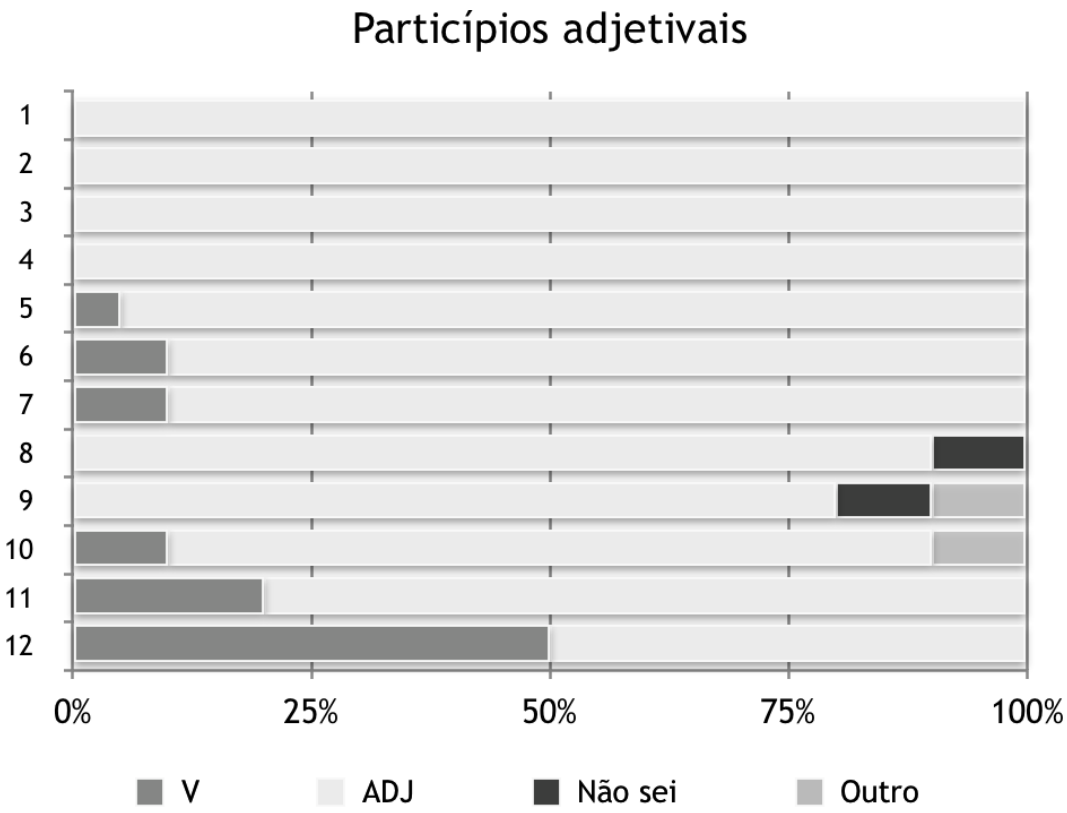

Gráfico 2 - Particípios adjetivais considerados fáceis

Frase 1 (100% ADJ): "Nefasta ideia, que deveria estar totalmente erradicada, à luz de estudos recentes, publicados em prestigiosas revistas **especializadas**."

Frase 2 (100% ADJ): "A família Kennedy pediu a William Manchester -- um **conhecido** jornalista, autor e historiador -- que escrevesse o que viria a ser a versão autorizada do assassinato de JFK."

Frase 3 (100% ADJ): "A vida era **fechada**, escura, estratificada."

Frase 4 (100% ADJ): "As novas tecnologias de material de fundição e a discussão sobre técnicas de gestão para sobreviver em um mercado

globalizado são os destaques do 7º Congresso de Fundição e da 6ª Feira de Fundidos, Insumos e Equipamentos, abertos ontem em São Paulo."

Frase 5 (95% ADJ, 5% V — 9 participantes marcaram ADJ, 1 participante marcou ADJ/V): "Saneamento do Estoril em Águas **agitadas**."

Frase 6 (90% ADJ, 10% V — 9 participantes marcaram ADJ, 1 participante marcou V): "Eu não estou **preocupado** em brilhar sozinho neste momento."

Frase 7 (90% ADJ, 10% V — 9 participantes marcaram ADJ, 1 participante marcou V): "No dia 3 de janeiro deste ano, uma tentativa de roubo a carro-forte deixou sete pessoas **feridas** e também um assaltante morto em Santo André."

Frase 8 (90% ADJ, 10% Não sei — 9 participantes marcaram ADJ, 1 participante marcou Não sei): "O corte do atacante Bettega, principal astro da equipe, também deixou o país **desconfiado.**"

Frase 9 (80% ADJ, 10% Não sei, 10% Outro — 9 participantes marcaram ADJ, 1 participante marcou Não sei, 1 participante marcou Outro): "Polanski andava meio **apagado**, fazendo o gênero bom-moço, há alguns anos."

Frase 10 (80% ADJ, 10% V, 10% Outro — 8 participantes marcaram ADJ, 1 participante marcou V, 1 participante marcou Outro): "Bronzeados, na faixa dos 20 aos 35 anos, chegam em carros **importados** e conversam sobre o último fim-de-semana no litoral norte."

Frase 11 (90% ADJ, 20% V — 8 participantes marcaram ADJ, 2 participantes marcaram V): "O espiritismo se baseia na crença da sobrevivência da alma e da existência de comunicação, por meio de mediunidade, entre vivos e mortos, entre os espíritos encarnados e **desencarnados**.",

Frase 12 (50% ADJ, 50% V — 5 participantes marcaram ADJ, 5 participantes marcaram V): "Ficariam resguardadas desta efetiva desvalorização da moeda **indexada** dos ricos apenas as cadernetas de poupança."

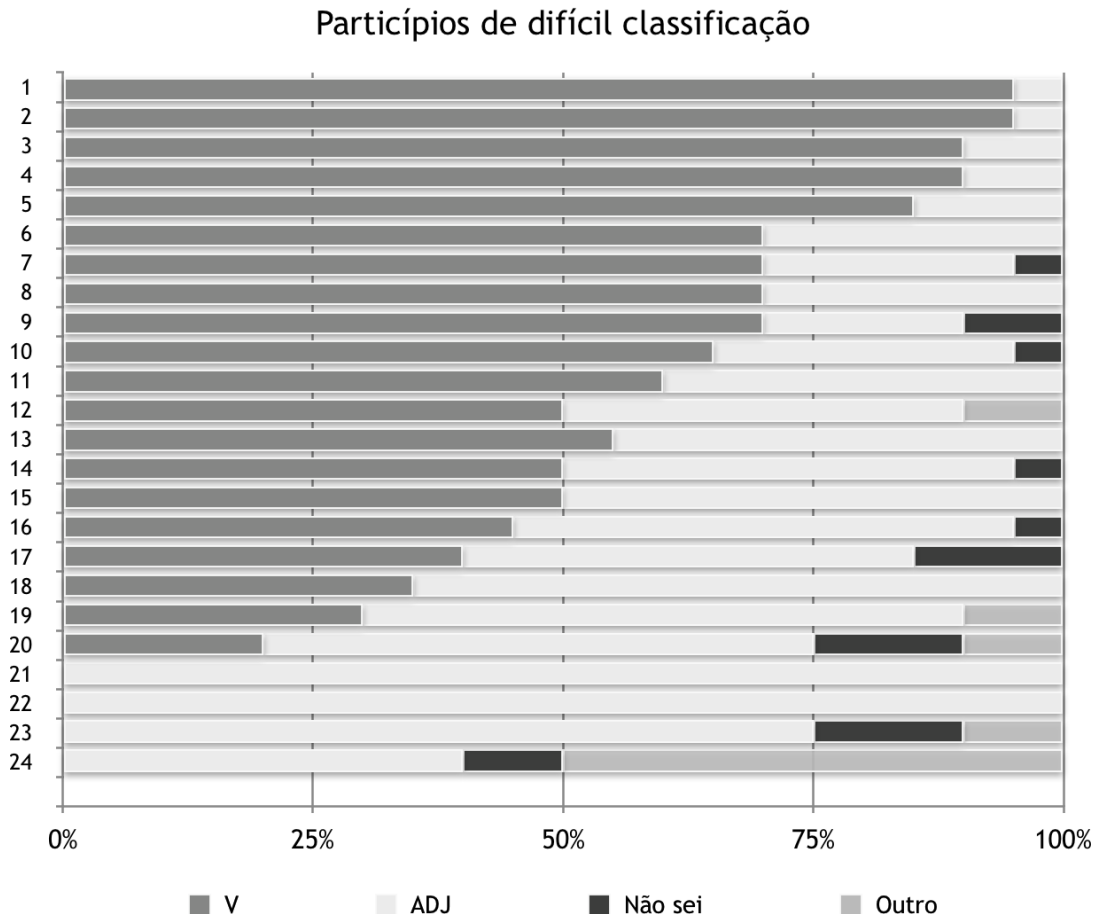

Gráfico 3: Particípios de difícil classificação

Frase 1: (95% V, 5% ADJ — 9 participantes marcaram V, 1 participante marcou V/ADJ): "É preciso varrer os cacos imprestáveis e limpar meticulosamente as peças **manchadas** pelo uso indevido."

Frase 2 (95% V, 5% ADJ — 9 participantes marcaram V, 1 participante marcou V/ADJ): "Ela depois me mandou um cartão agradecendo, e disse que o que mais tinha chamado a a atenção dela fora a palavra «ornament», que não era muito **usada**."

Frase 3 (90% V, 10% ADJ — 9 participantes marcaram V, 1 participante marcou ADJ): "Poucos foram no seu tempo tão **atacados**, poucos defenderam campanhas tão virulentas das forças conservadoras e em geral dos interessados em manter a rotina."

Frase 4 (90% V, 10% ADJ — 9 participantes marcaram V, 1 participante marcou ADJ): "A formação de engenheiros agrônomos até então era sistematicamente **orientada** para preparar profissionais no atendimento da demanda oferecida pelo governo e setor agroquímico."

Frase 5 (85% V, 15% ADJ) — 8 participantes marcaram V, 1 participante marcou ADJ, 1 participante marcou V/ADJ) — "Settimio Arturo Ferrazzetta, franciscano **nascido** em Verona (Itália) , em 1924, bispo da Guiné-Bissau desde 1977 (estava ali como missionário desde 1955) , começa por dizer que é «a pessoa menos indicada para falar» da Igreja do seu país de adopção, que ele apenas procura «servir da melhor maneira possível»."

Frase 6 (70% V, 30% ADJ) — 7 participantes marcaram ADJ, 3 participantes marcaram ADJ): "Ainda que se discuta o valor, o «peso» literário dos já **citados** e dos outros escritores do programa e da antologia -- afinal, que representatividade terão Domingos Pellegrini, Marina Colasanti e tantos outros? , o fato é que à Alemanha interessa outra coisa."

Frase 7 (70% V, 25% ADJ, 5% Não sei — 6 participantes marcaram V, 2 participantes marcaram ADJ, 1 participante marcou Não sei/V, 1 participante marcou ADJ/V): "O atual governador Hélio Garcia (PTB) e o governador **eleito** Eduardo Azeredo (PSDB) querem dois ministérios para o Estado."

Frase 8 (70% V, 30% ADJ — 7 participantes marcaram V, 3 participantes marcaram ADJ): "Por várias vezes **proibido** -- nomeadamente em alturas em que era preciso poupar pólvora, porque as lutas guerreiras consumiam todas as reservas -- o fogo de artifício nunca deixou de ser queimado."

Frase 9 (70% V, 20% ADJ, 10% Não sei — 7 participantes marcaram V, 2 participantes marcaram ADJ, 1 participante marcou Não sei): "Segundo a direção do Mário Gata, os outros detentos **feridos** durante a fuga também não correm risco de vida."

Frase 10 (65% V, 30% ADJ, 5% Não sei — 6 participantes marcaram V, 2 participantes marcaram ADJ, 1 participante marcou ADJ/Não sei, 1 participante marcou ADJ/V): "No primeiro caso, figuram o mexicano Roberto Sneider, diretor de «Dos Crimenes», o boliviano Juan Carlos Valdivia, de «Jonás y la Ballena Rosada», e o guatemalteco Luis Argueta, de «El Silencio de Neto», filmes bastante **aplaudidos** no Festival de Cartagena."

Frase 11 (60% V, 40% ADJ — 6 participantes marcaram V, 4 participantes marcaram ADJ): "Banana do Spot, um creme salgado de banana **temperado** com curry e outros condimentos (Cr$ 9.600,00)."

Frase 12 (50% V, 40% ADJ, 10% Outro — 5 participantes marcaram V, 4 participantes marcaram ADJ e 1 participante marcou Outro): "Refiro-me mesmo àqueles programas de interesse mais geral, como as telenovelas, que já nem essas entram às horas **anunciadas**."

Frase 13 (55% V, 45% ADJ — 5 participantes marcaram V, 4 participantes marcaram ADJ, 1 participante marcou ADJ/V) — "As pequenas e médias empresas nacionais, já **acostumadas** com a concorrência, sairão perdendo."

Frase 14 (50% V, 45% ADJ, 5% Não sei — 5 participantes marcaram V, 4 participantes marcaram ADJ, 1 participante marcou ADJ/Não sei): "Durante a festa, apareceu um cantor **vestido** de Elvis Presley."

Frase 15 (50% V, 50% ADJ — 5 participantes marcaram V, 4 participantes marcaram ADJ): "Enquanto se dirigia, apressada, a caminho da rádio difusora libertadores da América, que capitaneava uma rede interamericana **comprometida** com o desenvolvimento econômico do continente, Lina esforçava-se por não pensar em Edu, concentrando-se na sua missão."

Frase 16 (45% V, 50% ADJ, 5% Não sei — 4 participantes marcaram V, 5 participantes marcaram ADJ, 1 participante marcou Não sei/V): "12 fundos de alcachofras **cozidos**, 100 g de manteiga, 200 ml de creme de leite,

1 colher (sopa) de farinha de trigo, 4 colheres (sopa) de queijo ralado, sal, pimenta, gotas de limão."

Frase 17 (40% V, 45% ADJ, 15% Não sei — 4 participantes marcaram V, 4 participantes marcaram ADJ, 1 participante marcou Não sei, 1 participante marcou ADJ/Não sei): "É uma história **centrada** nos seus medos, «nas suas bravatas» (como diz a apresentação) , nas suas certezas «e, sobretudo, nas suas dúvidas»"

Frase 18 (30% V, 60% ADJ, 10% Outro — 3 participantes marcaram V, 5 participantes marcaram ADJ, 1 participante marcou Outro, 1 participante marcou Não sei): "Ele é **casado** com a ex-Ronaldinha Nédia França e vai ser papai em breve."

Frase 19 (20% V, 55% ADJ, 15% Não sei — 2 participantes marcaram V, 5 participantes marcaram ADJ, 1 participante marcou Não sei, 1 participante marcou Não sei/ADJ): "O incêndio que, desde anteontem, se regista em matas da freguesia de Chãs de Tavares, no concelho de Mangualde (que chegou a ser dado como **extinto** na manhã de ontem) , fez mesmo uma vítima mortal."

Frase 20 (75% ADJ, 25% V — 7 participantes marcaram ADJ, 2 participantes marcaram V, 1 participante marcou ADJ/V): "E não se incomodar em usar camisas que parecem **manchadas**."

Frase 21 (100% ADJ): "Boris SCHNAIDERMAN é ensaísta, crítico literário, tradutor e professor **aposentado** de língua e literatura russa da Universidade de São Paulo, e autor de A Guerra em Surdina."

Frase 22 (100% ADJ): "Ora, quando o Sul-Americano chegar ao fim, teremos no Rio e certamente em São Paulo a tão **desejada** temporada de Arsenal."

Frase 23 (75% ADJ, 15% Não sei, 10% Outro — 7 participantes marcaram ADJ, 1 participante marcou Não sei, 1 participante marcou Outro, 1 participante marcou ADJ/Não sei): "Pode ir desde encontrar um filme para assistir no cinema, até achar o endereço de um hospital em que um parente acaba de dar entrada e o telefone do famigerado lugar só dá **ocupado**."

Frase 24 (50% Outro, 40% ADJ, 10% Não sei — 5 participantes marcaram Outro, 4 participantes marcaram ADJ, 1 participante marcou Não sei): “Reich tinha uma forte orientação política, e não acreditava que as curas individuais poderiam acontecer em **separado** das principais mudanças sociais.”

Nossa classificação pessoal para essas 24 foi a seguinte: 1 - verbo, 2 - verbo, 3 - verbo, 4 - verbo, 5 - verbo, 6 - verbo, 7 - adjetivo, 8 - verbo, 9 - verbo, 10 - verbo, 11 - verbo, 12 - verbo, 13 - adjetivo, 14 - verbo, 15 - verbo, 16 - adjetivo, 17 - verbo, 18 - adjetivo, 19 - adjetivo, 20 - adjetivo, 21 - adjetivo, 22 - adjetivo, 23 - adjetivo - e 24 - substantivo.

Em três casos (7, 13 e 17), a maioria dos participantes divergiu da nossa classificação. Consideramos alterar nossa classificação para ficar de acordo com a opinião da maioria, mas optamos por não fazer isso porque a discrepância não era acentuada e porque não queríamos prejudicar a coerência de nossa anotação/conversão.

Os resultados do questionário com o Rêve apontam para a dificuldade de uma classificação consensual das formas participiais, o que não deixa de ser um argumento favorável à opção inicial do Mac-Morpho de manter uma classe PCP independente e da opção do Bosque de considerar todos os particípios como verbos (com a informação de serem também particípios). Adicionalmente, o fato de, originalmente, na classificação dos Estoicos, o particípio contar como uma forma gramatical independente (e não uma subclasse dos verbos) indica não ser essa uma opção sem qualquer motivação linguístico. Com isso em vista, decidimos criar um segundo dataset, com as classes do projeto UD e a classe PCP, a fim de verificar, posteriormente, se a manutenção da classe pode ser um facilitador na aprendizagem de pistas gramaticais.

5.3
Mac-Morpho com tagset UD + PCP

Na criação desse "novo" dataset, devolvemos ao corpus a etiqueta "PCP". Dessa forma, é possível investigar se manter essa etiqueta afeta o desempenho de um sistema e/ou afeta de forma significativa nas etapas posteriores à anotação de POS. Além disso, vale notar que o projeto UD conta com diferentes camadas para características adicionais. Assim, no futuro, pode-se manter a etiqueta V com a indicação "secundária" de ser um particípio (como é feito no Bosque).

Ao tomar essa decisão, porém, surgiram algumas questões: todos os particípios deveriam ser mantidos como PCPs ou apenas os que fossem ambíguos? Particípios em estruturas claramente verbais deveriam ser mantidos como PCPs? E particípios que apresentassem somente a leitura adjetival?

Optamos por manter todos os PCPs sem discriminação, pelos seguintes motivos: os particípios em estruturas verbais, por mais que estejam desempenhando o papel de verbos, ainda assim carregam traços adjetivais inexistentes em verbos comuns (como número e gênero) e, quanto aos adjetivos, não deve surpreender a constatação de que identificar particípios de leitura somente adjetival é uma tarefa extremamente difícil e controversa. Primeiramente, não há como identifica-los sem analisar o contexto, pois mesmo palavras já cristalizadas como adjetivos e que apresentam entrada nos dicionários como adjetivos (como "avermelhado" e "gelado", por exemplo) não apresentam necessariamente sempre essa leitura, podendo também aparecer em contextos de voz passiva com agente explícito. Por exemplo, em "sobremesas **geladas** para o verão", "geladas" atua como adjetivo, enquanto em "**gelados** pelo ar frio que os penetrava", "gelados" parece desempenhar um papel mais verbal; em "um outro pássaro, **avermelhado**, pousou […]", "avermelhado" é um adjetivo, já em "caras **avermelhadas** pela cachaça", apresenta características verbais. Além disso, considerar determinados particípios como adjetivos e outros não

inevitavelmente cria inconsistências e voltaríamos às classificações divergentes e remeteria à trabalhosa análise caso a caso que o Mac-Morpho procurou evitar ao instaurar a etiqueta PCP, e que nós também consideramos sábio evitar, sobretudo após os resultados do questionário. Com essa decisão, foram mantidos/re-inseridos formas participiais. Lembramos, no entanto, que não foi apenas uma "volta" ao Mac-Morpho inicial, uma vez que i) fomos mais rigorosas na manutenção do PCP, e mesmo casos considerados ADJ no Mac-Morpho foram transformados em PCP, como as palavras "adquirido" e "esperado"; b) conseguimos palavras que apareciam no Mac-Morpho etiquetadas como PCP, mas não deveriam estar, como as palavras "deputado" e "doutorado".

A seção seguinte trata das correções que efetuamos no Mac-Morpho.

6
Resultados: impactos da conversão e implicações para sistemas de PLN

Como mencionamos no capítulo anterior, uma das maneiras de verificar o impacto de diferentes tagsets em sistemas de PLN é comparando o desempenho de um (ou diferentes) sistema(s) com o mesmo material, com alteração apenas no tagset.

Outra maneira de avaliar tagsets de POS consiste em avaliar o impacto do tagset em tarefas subsequentes, como na análise de dependências. A segunda maneira nos parece mais interessante, mas a ausência de dependências no Mac-Morpho torna, por enquanto, esta tarefa impossível de ser realizada. O que fizemos, portanto, foi utilizar o sistema — uma rede neural recursiva de múltiplas camadas, ainda em desenvolvimento na data de publicação — de Rafael Rocha, da equipe do LEARN (PUC-Rio), para a anotação no corpus Mac-Morpho, utilizando os 3 seguintes datasets:

a) Mac-Morpho com tagset UD;

b) Mac-Morpho com tagset UD + PCP;

c) Mac-Morpho "clássico" (com tagset Mac-Morpho);

O quadro abaixo apresenta a acurácia do sistema, conforme o dataset utilizado:

Mac-Morpho com tagset UD + PCP	98,17%
Mac-Morpho com tagset UD	98,12%
Mac-Morpho com tagset Mac-Morpho	97,94%

Quadro 5: resultados do sistema de Rocha (2016) com os diferentes datasets.

Na avaliação, o sistema alcançou 97.94% para o corpus Mac-Morpho v. 1 com tagset do próprio Mac-Morpho, 98.12% para o corpus Mac-Morpho v. 1 com o tagset UD e 98.17% para o corpus Mac-Morpho v.

1 com o tagset UD acrescido da etiqueta PCP. Como era de se esperar, o desempenho com o tagset UD foi ligeiramente melhor que com o tagset do Mac-Morpho, uma vez que seu dataset é menos granular. Por sua vez, o desempenho com o tagset UD + PCP foi ainda um pouco melhor.

Para verificar se houve algum impacto relativo à mudança no tagset, fizemos uma breve análise dos erros, tendo por base a matriz de confusão de cada saída e a análise de uma amostra das frases que contêm elementos errados.

6.1
Mac-Morpho com tagset Mac-Morpho

Consultando a matriz de confusão para o desempenho com corpus Mac-Morpho v. 1 com o tagset do próprio Mac-Morpho, a maior quantidade de erros (acima de 150 ocorrências) está localizada em[40]: N > NPROP (507 erros), NPROP > N (452 erros), ADJ > N (281 erros), N > ADJ (186 erros) e VAUX > V (158 erros).

A maior quantidade de erros foi gerada pela confusão de N com NPROP e vice-versa, que é um erro compreensível e esperado. Afinal, nomes próprios são uma subcategoria de nomes/substantivos, e muitas vezes a distinção entre ambos pode ser sutil, dificultando o aprendizado de máquina.

É interessante destacar também que, em alguns casos, os erros são causados por conta de escolhas na etapa de tokenização. Por exemplo, todas as ocorrências de palavras que foram consideradas preposição quando a etiqueta original era de nome próprio (NPROP > PREP, no caso, 69 erros) são referentes a palavras que de fato são preposições, mas faziam parte de alguma expressão que foi por inteiro etiquetada como nome próprio, como no caso de:

[40] O primeiro elemento é a etiqueta original e o segundo, após o sinal de >, a etiqueta predita.

(80) “Pedro Camargo Neto, ex-presidente de a SRB, lançou ontem, em São Paulo, o livro "Pensamento Rural reflexões sobre{NPROP>PREP} o{NPROP>ART} desenvolvimento{NPROP > N} brasileiro{NPROP > ADJ}".”

Caso as concatenações de etiquetas do Mac-Morpho fossem feitas de maneira explícita, como é referido no manual, ou desfeitas e cada palavra recebesse uma etiqueta referente a si própria, é possível que o desempenho do sistema melhorasse.

6.2
Mac-Morpho com tagset UD

A matriz de confusão para o desempenho com corpus Mac-Morpho v. 1 com o tagset UD informa que a maior quantidade de erros (acima de 150) concentrou-se em NOUN > PROPN (463 erros), PROPN > NOUN (392 erros), NOUN > ADJ (304 erros), ADJ > VERB (251 erros), AUX > VERB (236 erros), VERB > ADJ (194 erros), NOUN > VERB (176 erros) e ADJ > NOUN (175 erros).

Ao comparar os resultados do Mac-Morpho original com o Mac-Morpho com tagset UD, percebemos que a conversão para um tagset sem a etiqueta “PCP” gerou um grande aumento na confusão entre adjetivos e verbos (em ambas as direções), confusão praticamente inexistente no original, mas mesmo assim o desempenho do sistema com esse dataset foi superior ao utilizando o original. Essa melhora no desempenho não chega a ser surpreendente, uma vez que o tagset UD é bem menos granular do que o do Mac-Morpho — como o tagset UD apresenta menos etiquetas, o sistema tem menos chances de errar e como os erros espalhados por outras classes, apesar de poucos, deixaram de existir, o impacto no resultado final foi visível.

Decidimos, então, com o dataset UD + PCP, nosso principal interesse, conduzir uma análise mais detalhada dos erros. Para analisar os tipos de erros, consideramos 10% do total de erros de cada categoria e

analisamos as frases individualmente. Não pretendemos com isso fazer uma análise extensiva dos erros, apenas conduzir um pequeno estudo exploratório para evidenciar o potencial desse tipo de análise, como já realizado em Freitas et al. (2006) e Manning (2011).

Listamos as análises a seguir, distribuídas por pares (quando houver pares) de confusão:

i) Confusão entre substantivo e nome próprio, em ambas as direções (NOUN > PROPN e PROPN > NOUN)

Confusão entre substantivo e nome próprio (ambos os sentidos) Frases analisadas: 85 Tipo de erro:
Erros no golden: 28
Erros “aleatórios”: 38
Tokenização: 19

Quadro 6: confusão entre substantivo e nome próprio — tipos de erro.

A maior parte da confusão entre substantivo e nome próprio pode ser atribuída ao que chamamos de “erros” aleatórios. São estes os erros do sistema sem padrão ou cujo padrão não conseguimos identificar por conta da amostra reduzida, casos em que não há motivo aparente para o sistema errar, mas nos quais houve erro mesmo assim, indicando que não aprendeu a distinguir as classes corretamente. Exemplificamos esse fenômeno com a seguinte frase:

(81) “As duas sagas (narrativas tradicionais islandesas) de esse volume contam como Érik{PROPN > NOUN}, o Vermelho e seu filho, Leif, o Venturoso, realizaram a proeza”

Pode-se ver também que há um número razoável do que chamamos de erros do golden. Nesses casos, os “erros” são palavras que o sistema na verdade etiquetou corretamente, mas que apareciam no golden (que funciona como um gabarito) anotadas com uma etiqueta errada, de forma que o acerto do sistema é apontado como erro. Por exemplo:

(82) “O prefeito de Campinas{NOUN > PROPN} (SP), José Roberto Magalhães Teixeira (PSDB), visitou ontem a Folha, onde foi recebido em almoço.”

Outro grande motivo para a confusão pode ser devido à opção de tokenização do Mac-Morpho de manter as palavras de uma expressão multivocabular como tokens separados, porém com a mesma etiqueta (o que, como já dissemos anteriormente, destoa da opção original, documentada no manual do Mac-Morpho). Devido a essa decisão, palavras que seriam normalmente anotadas como substantivos aparecem como nome próprio por estarem fazendo parte do “nome” de algo, como se pode ver no exemplo abaixo:

(83) “"Se deixar , vai longe" , diz Meneguello , rindo , antes de introduzir o novo tema : O{PROPN > PRON} que{PROPN > PRON} é{PROPN > VERB} o{PROPN > DET} medo{PROPN > NOUN}?”

ii) Confusão entre substantivo e adjetivo, em ambas as direções (NOUN > ADJ e ADJ > NOUN)

Confusão entre substantivo e adjetivo (ambos os sentidos) Frases analisadas: 47 Tipo de erro:
Erros no golden: 13
Erros "aleatórios": 11
Flutuação NOUN/ADJ: 20
Tokenização: 3

Quadro 7: confusão entre substantivo e adjetivo — tipos de erro.

Dois grandes motivos das confusões NOUN > ADJ e ADJ > NOUN são erros no golden e o não aprendizado completo da distinção entre as duas classes (causando os erros “aleatórios”). Porém, a maior fonte de confusão entre essas duas classes é o que chamamos de “flutuação NOUN/ADJ”: a ocorrência de palavras que podem funcionar tanto como adjetivo quanto

como substantivo, aparecendo com ambas as anotações no corpus e confundindo assim o sistema, como nos exemplos:

(84) “A os jovens{ADJ > NOUN} gregos{NOUN > ADJ} destinava se não a "didaskalia " ensinamento de uma profissão - mas a "paidéia", a formação ("Bildung")”;

(85) “Ao lado de poucos ricos convivem muitos{DET > PRON} pobres{NOUN > ADJ} , com graves problemas nutricionais”;

(86) “Campeonatos italiano, inglês{ADJ > NOUN}, espanhol{ADJ > NOUN}, japonês{ADJ > NOUN}, holandês{ADJ > NOUN}”.

iii) Confusão entre verbo auxiliar e verbo (AUX > VERB)

Confusão entre verbo auxiliar e verbo Frases analisadas: 23 Tipo de erro:
Erros no golden: 4
Erros "aleatórios": 14
Distanciamento entre AUX e VERB: 5

Quadro 8: confusão entre verbo auxiliar e verbo — tipos de erro.

A maior parte dos erros dessa confusão deve-se ao fato de o sistema aparentemente não ter aprendido totalmente a distinção entre verbo auxiliar e verbo. Assim, quando há um verbo auxiliar seguido de um verbo principal, por vezes o sistema identifica as duas palavras como verbos (uma interpretação/análise que até nos parece possível, mas não é o padrão de anotação do Mac-Morpho), como os exemplos a seguir evidenciam:

(87) “O Instituto Ecoar para a Cidadania está{AUX > VERB} doando mudas de eucaliptos para pequenos e médios produtores rurais”

(88) “A indústria brasileira de curtume foi{AUX > VERB} buscar know-how em a Europa para melhorar a qualidade de o couro nacional .”

(89) “O próprio ministro de a Fazenda{PROPN > NOUN} , Fernando Henrique Cardoso , e o secretário executivo , Clóvis Carvalho , já

estavam{AUX > VERB} isolando Fritsch de as principais decisões de a área econômica”

Os outros erros que encontramos deviam-se a erros no golden e ao distanciamento entre o verbo auxiliar e o verbo principal, o que tem potencial para confundir a aprendizagem do sistema, conforme demonstra o seguinte exemplo:

(90) “É outra alternativa que deve ser{AUX > VERB} melhor explorada por os produtores " , diz Denardin”.

iv) Confusão entre verbo e adjetivo, em ambas as direções (VERB > ADJ e ADJ > VERB)

Confusão entre verbo e adjetivo (em ambas as direções) Frases analisadas: 44 Tipo de erro:
Formas participiais: 44

Quadro 9: confusão entre verbo e adjetivo — tipos de erro.

Como os quadros acima evidenciam, das frases que analisamos, todos os erros deviam-se às formas participiais. Como o papel dos particípios por vezes depende de um contexto amplo e é de difícil distinção até mesmo para humanos, era de se esperar que a o desempenho do sistema não fosse perfeito. A seguir, exemplos desse tipo de erro:

(91) “Dez promotores de o Rio e agentes de o serviço reservado{ADJ > VERB} de a polícia militar apreenderam em a última quarta-feira documentos com nomes de políticos e policiais que receberiam propinas de bicheiros”;

(92) “Ela foi sequestrada e seu corpo encontrado{VERB > ADJ} , com oito tiros , em um lixão em Parati (RJ)”

v) Confusão entre substantivo e verbo (NOUN > VERB):

Confusão entre substantivo e verbo Frases analisadas: 17 Tipo de erro:
Erros no golden: 5
Erros "aleatórios": 1
Terminação verbal/participial: 7
Forma participial usada como NOUN: 4

Quadro 10: confusão entre substantivo e verbo — tipos de erro.

O sistema parece confundir-se com formas nominais idênticas a verbos/particípios ou com terminações caracteristicamente verbais, como nos exemplos:

(93) “Feder afirmou que vai voltar a o PL, partido{NOUN > VERB} a o qual ele se filiou em 1986”

(94) “Arte barroca, selos e tapeçaria{NOUN > VERB}”, respectivamente”

Pode-se ver também que houve um bom número de erros causados por erros no golden e em casos em que um particípio aparecia atuando como substantivo, como no exemplo:

(95) “As arbitragens fraudulentas em o futebol fluminense - o " jogo roubado por o juiz " motivaram até confissões de envolvidos{NOUN > VERB} , mas o esquema que as montou não poderá ser investigado por CPI”

6.3
Mac-Morpho com tagset UD + PCP

Por fim, consultando a matriz de confusão para o desempenho com corpus Mac-Morpho v. 1 com o tagset UD acrescido da etiqueta “particípio”, a maior quantidade de erros foi NOUN > PROPN (605), PROPN > NOUN (325), AUX > VERB (284), NOUN > ADJ (266) e ADJ > NOUN (186). É interessante notar que a confusão entre verbos e adjetivos

(em ambas as direções) desapareceu da lista. Porém, alguns valores aumentaram, como os erros NOUN > PROPN, mas não sabemos o que pode ter motivado essa queda de desempenho em comparação ao Mac-Morpho com tagset UD.

7
Conclusões e considerações finais

A principal motivação deste trabalho foi propiciar um cenário que viabilizasse o estudo do impacto de diferentes tagsets em sistemas de PLN. Ao longo desta dissertação, apresentamos uma proposta de conversão entre tagsets sem etiquetas diretamente equivalentes e, com isso, reanotamos o corpus Mac-Morpho. Partindo dessa motivação, originalmente vinda da área de PLN, conduzimos um estudo linguístico sobre particípios passados, categoria que se revelou nosso maior desafio de conversão. Ao fim da pesquisa havíamos criado 3 datasets[41], ainda que a ideia original fosse a criação de apenas um: além do Mac-Morpho alinhado com o tagset UD, criamos uma versão que adiciona a etiqueta PCP à versão UD. Além disso, ao longo do processo de elaboração das regras de conversão, encontramos erros de anotação no Mac-Morpho e também criamos regras para corrigi-los. Assim, aplicando apenas essas regras, pudemos obter uma nova versão revista do Mac-Morpho.

Já de antemão, sabíamos que haveria nesta empreitada espaço para estudos linguísticos devido à presença, no Mac-Morpho, de uma etiqueta de forma participial, que precisaria ser convertida em verbo ou adjetivo. Nesse sentido, a tarefa de conversão apresentou-se como uma motivação para estudar essa "classe" reconhecidamente complicada, mas pouco estudada, sobretudo com base em grandes corpora. Acreditamos que, com a dissertação, contribuímos com mais dados e análise com relação aos particípios, como ilustra o material do capítulo 5.

Em relação ao trabalho de conversão de tagsets, é relevante frisar que existem sempre duas etapas: o alinhamento entre as etiquetas e entre as filosofias gramaticais subjacentes à aplicação dessas mesmas etiquetas. Mesmo quando inicialmente parece haver um alinhamento entre as etiquetas, é de extrema importância – além de ler a documentação

[41] Todos estão disponíveis em https://github.com/own-pt/macmorpho-UD

cuidadosamente – analisar uma amostra considerável de exemplos, pois frequentemente existem decisões linguísticas não explicitadas (o caso das formas participiais é uma boa ilustração). Muitas vezes, não apenas nuances de classificação podem resultar em um alinhamento diferente do que parecia funcionar na teoria, como as filosofias de anotação e tokenização dos corpora de cada tagset podem ser diferentes, o que pode acarretar na etiquetagem equivocada, imprecisa ou indesejada de alguns elementos, impactando negativamente na consistência do recurso e, consequentemente, no desempenho de sistemas que aprendem por meio de exemplo e na confiança da avaliação que toma por base o recurso (o golden). Ao longo desta pesquisa, tornou-se evidente para nós que não é possível atingir uma conversão de qualidade realizando o processo de forma puramente automática.

Consideramos importante ressaltar que procuramos não apenas converter o tagset, mas também padronizar a filosofia de anotação e as decisões gramaticais de acordo com os manuais do UD (por exemplo, eliminar a concatenação de etiquetas). Porém, como o Mac-Morpho é um corpus bastante grande e, devido à falta de tempo e de mão de obra, não foi possível (re)revisá-lo inteiro manualmente (apesar de termos revisado partes específicas afetadas por nossas regras e amostras aleatórias do corpus), é muito provável que ainda haja elementos no padrão Mac-Morpho "clássico".

Outra contribuição deste trabalho é a proposta de que se tome como base a matriz de confusão no processo de revisão do golden, pois isso permite uma revisão mais precisa e direcionada, o que, por sua vez, possibilita o aprimoramento do recurso de forma mais eficaz.

Como a identificação das diferentes filosofias de anotação para o alinhamento adequado tomou mais tempo que o esperado, e não foi possível alinhar todos os pontos, de forma que ficam por fazer: a padronização dos

verbos modais para o modelo UD[42] e a desconcatenação das etiquetas concatenadas restantes (com a adição da informação de que se tratam de expressões multivocabulares em outro nível de anotação, conforme explicamos no capítulo 3).

Em relação às classes de palavras, este trabalho nos permitiu perceber que que muitas vezes as classes tradicionalmente apresentadas nas gramáticas podem não ser as ideais para tratar de certos fenômenos linguísticos no contexto aplicado das tarefas do PLN. Conforme já foi dito ao longo deste trabalho, as classes de palavras atualmente utilizadas na classificação do português são fruto de escolhas humanas, com interesses pautados em contextos histórico-sociais diferentes do atual. Assim, é importante reconhecer que justamente por isso é possível questioná-las e repensá-las. Esse fato também oferece uma justificativa linguisticamente motivada para o investimento em discussões e propostas de outras classificações.

Em relação aos particípios, a pesquisa em corpus trouxe informações para lidar com o fenômeno que até então eram indisponíveis, o que é extremamente frutífero para a área de descrição gramatical. A literatura sobre o tema apresenta visões e pontos válidos, porém em geral baseados em intuições e exemplos controlados, sem levar em consideração (por falta de meios para fazê-lo) como têm sido usadas as formas participiais: em que contextos costumam aparecer, como seus papéis mudam de acordo com os elementos que os cercam, como variam de acordo com o gênero e o registro linguísticos, dentre outras informações relevantes para uma abordagem mais rica e abrangente do fenômeno. Consideramos relevante ressaltar que, atualmente, é possível realizar o estudo de fenômenos linguísticos levando em conta grandes corpora, o que contribui para retratos mais abrangentes de certos fenômenos da língua.

---

[42] O padrão do UD é etiquetar verbos modais como verbos auxiliares, enquanto a do Mac-Morpho é anotá-los como verbos. Não efetuamos as alterações dos modais nessa primeira conversão porque essa divergência só foi descoberta na última etapa do trabalho, mas isso certamente será feito em aprimoramentos posteriores. Trata-se de mais um exemplo ligado a diferentes concepções gramaticais — ou filosofias de anotação, como chamamos genericamente neste trabalho.

Em especial, ainda com relação aos particípios, chamamos a atenção para dois pontos: i) a dificuldade de consenso das análises, informada pelos resultados do questionário com o Rêve, e ii) a relevância da classificação ou distribuição dessas formas para sistemas de PLN, que pode ser significativa ou não.

Levantamos o segundo ponto após realizar uma análise preliminar, com 30 frases contendo formas participiais presentes no corpus UD-PT e no Bosque (que sempre considera os particípios como verbos) e analisar o papel desempenhado pelo particípio nas dependências sintáticas e os elementos de quem o particípio é "filho". Para nossa surpresa, averiguamos que, independentemente do particípio ser tratado como adjetivo em um corpus e como verbo em outro, as dependências mantinham-se iguais nos dois corpora, na grande maioria dos casos. Ou seja, manter a etiqueta "particípio" pode ter um impacto muito menor nas dependências do que supúnhamos a princípio. Para outras aplicações, como métricas de complexidade textual e papeis semânticos, essa distinção dos particípios pode parecer mais significativa, quiçá essencial. Entretanto, essa impressão se deve ao fato da maioria esmagadora da área de anotação de POS basear-se em um modelo gramatical específico e (como vimos) contestável, o que cria a necessidade artificial de encaixar todos fenômenos linguísticos em algumas classes limitadas. Repensando a abordagem, porém, é possível que seja mais proveitoso manter particípios classificados como tal, ao invés de forçadamente separá-los em adjetivos ou verbos (ou substantivos ou advérbios). Consideramos fundamental verificar futuramente a real relevância de distinguir particípios entre verbos e adjetivos, por meio da observação de resultados de outros contextos (como estudos mais direcionados no âmbito das dependências, de papeis semânticos etc).

Conforme o previsto, ao longo desta pesquisa conseguimos criar um cenário propício para a verificação do impacto de diferentes tagsets de POS em tarefas de PLN de língua portuguesa. Por meio do acesso a um mesmo corpus anotado com diferentes tagsets e a um mesmo sistema, foi possível

observar o impacto de diferentes tagsets no desempenho do sistema. Foi nesse sentido que conduzimos, no capítulo 6, um breve estudo sobre o impacto das conversões, mas sabemos tratar-se apenas de uma exploração preliminar. Não temos ainda como verificar o impacto dos tagsets em etapas posteriores do processamento, mas trata-se de um estudo relevante no qual temos grande interesse para o futuro.

Os resultados que obtivemos em relação ao desempenho do sistema e às análises de erro preliminares que conduzimos indicam que a acurácia do sistema não é alterada de forma tão significante quando se utiliza um tagset pensado especificamente para a língua do corpus ou um criado com o objetivo de ser universal. Como esperado, o sistema teve um desempenho melhor utilizando o corpus anotado com o tagset com menor número de etiquetas (que era, no caso desta pesquisa, o universal), mas não foi uma diferença de grande magnitude.

Após a análise dos erros, pudemos concluir que, em muitas ocasiões, boa parte dos erros são, na realidade, predições corretas do sistema que apenas aparecem como erros porque a anotação do golden está equivocada (Manning, 2011, atribui 15,5% dos erros do sistema a erros do golden e 28% a inconsistências/ausência de padrão no golden). A análise de erros, portanto, é uma boa forma de efetuar uma revisão mais direcionada do golden, melhorando a qualidade do recurso.

Dessa forma, os resultados, em última análise, apontam para o caráter fundamental do investimento na qualidade dos recursos. Frequentemente procura-se corrigir o desempenho de um sistema utilizando como base dados que podem estar comprometidos por erros no golden, o que acaba levando, na verdade, à piora do sistema – ainda que a uma melhoria na quantidade de acertos..

Outro dado interessante que percebemos foi que, com alguma frequência, os erros se deviam a pontos realmente dependentes de análises gramaticais e/ou de classes cuja flutuação entre os membros é reconhecida, como a flutuação entre N e ADJ e as análises relativas às locuções verbais.

Algo que consideramos ter um potencial interessante e que gostaríamos de testar caso houvesse mais tempo é a criação de uma versão do corpus Mac-Morpho sem etiquetas concatenadas, com etiquetas referentes apenas ao token ao qual estão acopladas, independentemente destes fazerem ou não parte de uma expressão multivocabular. Após realizar uma análise de erros detalhada do desempenho de um sistema anotador utilizando o corpus Mac-Morpho com diferentes tagsets, percebemos que a maioria dos erros gerados na anotação automática do corpus Mac-Morpho com o tagset Mac-Morpho era devido à concatenação de etiquetas[43]. Ao que parece, removendo a concatenação (mas ainda indicando a existência de MWEs em outro nível de anotação, de forma que essa informação não seja perdida), o desempenho para esse corpus com esse tagset aumentaria, possivelmente superando o desempenho obtido com os outros tagsets.

Para o futuro, também pretendemos "enxugar" as nossa lista de regras criadas, principalmente visando remover redundâncias e, sobretudo, identificar padrões, de forma a reduzir o número de regras e torná-las mais otimizadas e eficazes.

Duas etapas que consideramos extremamente interessantes e relevantes, mas que desde o início soubemos que não teríamos tempo de executar, são a lematização do Mac-Morpho e o estabelecimento de dependências. A implementação dessas etapas deixaria o corpus em maior sintonia com os padrões do projeto UD, o tornaria ainda mais útil para os usuários e abriria uma enorme gama de possibilidades de pesquisas.

Por fim, esperamos ter demonstrado, com este trabalho, como pode ser proveitoso o diálogo entre a descrição de uma língua e a Linguística Computacional.

---

43 Já temos conhecimento de que existe também uma versão do Mac-Morpho com as etiquetas formalmente concatenadas, mas não há tempo de produzir novos dados para esta dissertação.

Referências bibliográficas

AFONSO, S., BICK, E., HABER, R., & SANTOS, D. Floresta sintá(c)tica: a treebank for Portuguese. In Proceedings of the Third International Conference on Language Resources and Evaluation. LREC, 2002. p. 1698–1703.

ALUÍSIO, S., PELIZZONI, J., MARCHI, A. R., DE OLIVEIRA, L., MANENTI, R., & MARQUIAFÁVEL, V. An account of the challenge of tagging a reference corpus for brazilian portuguese. In: Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language. PROPOR, 2003.

ALUÍSIO, S. Anotação de Corpus: uma área de integração entre linguistas e linguistas computacionais. Apresentação na mesa redonda 10 anos do ELC: a LC no Brasil. X Encontro de Linguística de Corpus, 2011.

AUROUX, S. A revolução tecnológica da gramatização. Campinas, SP, Editora da Unicamp, 1992.

BAGNO, M. Gramática pedagógica do português brasileiro. Editora Parábola, 2012.

BASILIO, M. Formação e classes de palavras no português Brasil. Editora Contexto, 2008.

BECHARA, E. Moderna gramática portuguesa. 11 ed. São Paulo: Companhia Editora Nacional, 1967.

BIDERMAN, M. T. C. Teoria lingüística: teoria lexical e lingüística computacional. Martins Fontes, 2001.

CUNHA, C. & CINTRA, L. Nova gramática do português contemporâneo. Vol. 2. Rio de Janeiro: Nova Fronteira, 2001.

CÂMARA, J. M. Dicionário de filologia e gramática, referente à língua portuguêsa. J. Ozon, 1970.

DÉJEAN, H.. How To Evaluate and Compare Tagsets? A Proposal. LREC, 2000.

FOLTRAN, M. J. & CRISÓSTIMO, G.. Os adjetivos participiais no português. Revista de Estudos da Linguagem, 2005. 13.1: p. 129-154.

FONSECA, E. R. & ROSA, J. L. G. Mac-Morpho revisited: Towards robust part-of-speech tagging. In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, 2013. p. 98-107.

FONSECA, E. R., ROSA, J. L. G., ALUÍSIO, S. M. Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. Journal of the Brazilian Computer Society, 2015.

FREITAS, C. Esqueleto – investigação sobre o léxico do corpo para a inclusão de informação semântica em corpora da língua portuguesa. Projeto de Bolsa de Produtividade aprovado pela PUC-Rio, 2013.

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, 2015. p 18-46. (PDF)

FREITAS, C. & AFONSO, S. Bíblia Florestal: Um manual lingüístico da Floresta Sintá (c) tica, 2007.

GARSIDE, R., LEECH, G. N., MCENERY, T. (ed.). Corpus annotation: linguistic information from computer text corpora. Taylor & Francis, 1997.

ILARI, R. & BASSO, R. O português da gente: a língua que estudamos: a língua que falamos. Editora Contexto, 2006.

KILGARRIFF, Adam; KOSEM, Iztok. Corpus tools for lexicographers. na, 2012.

LEECH, G. & WILSON, A. EAGLES recommendations for the morphosyntactic annotation of corpora. Version of March, 1996.

LEECH, G. Adding linguistic annotation. In: Developing linguistic corpora : a guide to good practice. Oxbow Books, Oxford, 2005. p. 17-29.

MACAMBIRA, J.R. A estrutura sintática do português. 5 ed. São Paulo: Livraria Pioneira, 1987.

MANNING, C. D.; SCHÜTZE, Hinrich. Foundations of statistical natural language processing. Cambridge: MIT press, 1999.

MANNING, C. D. Part-of-speech tagging from 97% to 100%: is it time for some linguistics?. In: International Conference on Intelligent Text Processing and Computational Linguistics. Springer Berlin Heidelberg, 2011. p. 171-189.

MITKOV, R. The Oxford handbook of computational linguistics. Oxford University Press, 2005.

NEVES, M. H. M. A vertente grega da gramática tradicional: uma visão do pensamento grego sobre a linguagem. Editora UNESP, 2005.

NUNES, M. G. V., GHIRALDELO, C. M., MONTILHA, G., TURINE, M. A. S., DE OLIVEIRA, M. C. F., HASEGAWA, R. & OLIVEIRA JR,

O. N. . Desenvolvimento de um sistema de revisão gramatical automática para o português do Brasil. In: Anais do II Encontro para o processamento de português escrito e Falado. Curitiba: CEFET-PR, 1996. p. 71-80.

OLIVEIRA, C. & FREITAS, M. C. Classes de palavras e etiquetagem na Lingüística Computacional. Calidoscópio, 2006. 4.3: 179-188.

PEREIRA, M.T. Palavras denotativas: temas e problemas. In: Flores Verbais, Uma Homenagem Lingüística e Literária para Eneida do Rego Monteiro Bomfim no seu 70° Aniversário. Heye, J. (org). Rio de Janeiro: 34 Editora, 1995. p. 15-21.

PERINI, M. A. Sofrendo a Gramática. Ed. Ática, São Paulo, 1997.

PETROV, S., DAS, D., & MCDONALD, R.. A universal part-of-speech tagset, 2011.

PIMENTA-BUENO, M.. As formas V+do do português: um estudo de classes de palavras. DELTA, 1986. 2(2) p. 207-229.

RAMISCH, C. A generic framework for multiword expressions treatment: from acquisition to applications. In: Proceedings of ACL 2012 Student Research Workshop. Association for Computational Linguistics, 2012. p. 61-66.

ROSA, M. C. Introdução à morfologia. Editora Contexto, 2000.

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

SANTOS, D. Gramateca: corpus-based grammar of Portuguese. In: BAPTISTA, J., MAMEDE, N., CANDEIAS, S., PARABONI, I., PARDO, T. A. S. & NUNES, M. G. V. (eds.), International Conference on Computational Processing of Portuguese (PROPOR'2014), São Carlos, 2014. Springer, pp. 214—219. http://www.linguateca.pt/Diana/download/gramateca.pdf

SANTOS, D., MARQUES, R., FREITAS, C., SIMÕES, A., & MOTA, C.. Comparando anotações linguísticas na Gramateca: filosofia, ferramentas e exemplos. Domínios de Lingu@gem, 2015. v. 9, p. 11-26.