UNIVERSIDADE ESTADUAL DE CAMPINAS
Instituto de Estudos da Linguagem

IGOR LEAL SOUZA

A INFORMATIVIDADE DA MORFOLOGIA PARA A CATEGORIZAÇÃO DISTRIBUCIONAL DE PALAVRAS: UM MODELO COMPUTACIONAL

Campinas
2022

IGOR LEAL SOUZA

A INFORMATIVIDADE DA MORFOLOGIA PARA A CATEGORIZAÇÃO DISTRIBUCIONAL DE PALAVRAS: UM MODELO COMPUTACIONAL

Dissertação apresentada ao Instituto de Estudos da Linguagem da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Linguística.

Orientador: Prof. Dr. Pablo Picasso Feliciano de Faria

Este exemplar corresponde à versão final da dissertação defendida pelo aluno Igor Leal Souza e orientada pelo Prof. Dr. Pablo Picasso Feliciano de Faria.

Campinas
2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Estudos da Linguagem
Tiago Pereira Nocera - CRB 8/10468

L473i

Leal, Igor, 1991-
A informatividade da morfologia para a categorização distribucional de palavras : um modelo computacional / Igor Leal Souza. – Campinas, SP : [s.n.], 2022.

Orientador: Pablo Picasso Feliciano de Faria.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Estudos da Linguagem.

1. Processamento de linguagem natural (Computação). 2. Processamento de dados. 3. Análise por agrupamento. 4. Morfologia. 5. Aquisição da linguagem. I. Faria, Pablo, 1978-. II. Universidade Estadual de Campinas. Instituto de Estudos da Linguagem. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: The informativeness of morphology for the distributional categorization of words : a computational model
Palavras-chave em inglês:
Natural language processing (Computer science)
Data processing
Cluster analysis
Morphology
Language aquisition
Área de concentração: Linguística
Titulação: Mestre em Linguística
Banca examinadora:
Pablo Picasso Feliciano Faria
Marcelo Barra Ferreira
Maria Cristina Lobo Name
Data de defesa: 19-04-2022
Programa de Pós-Graduação: Linguística

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-3022-895X
- Currículo Lattes do autor: http://lattes.cnpq.br/5523830537982736

BANCA EXAMINADORA

Pablo Picasso Feliciano de Faria

Marcelo Barra Ferreira

Maria Cristina Lobo Name

IEL/UNICAMP
2022

Ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-Graduação do IEL.

Dedico este trabalho a todas as pessoas que me apoiaram nessa caminhada, em especial, ao meu tio Adauto (*in memoriam*), sem o qual eu não conseguiria chegar até aqui.

AGRADECIMENTOS

Não sei se toda jornada precisa de um fim, mas essa, com certeza, é uma das que precisava. Foram muitas experiências, boas e ruins, que jamais imaginei passar. E, nesse período, uma pessoa sempre ficou ao meu lado e embarcou comigo em todas as decisões, me dando força e coragem para prosseguir, foi Priscila Tuy, companheira para todas as horas e que, sem hesitar, aceitou a proposta de abrir mão de tudo que construímos para iniciar uma nova vida em Campinas, sem saber o que aconteceria. Com certeza, nunca imaginamos que uma pandemia iria acontecer (não mesmo!) e que durante essa passagem tão difícil, ainda conseguiríamos nos apoiar e superar as dificuldades apresentadas pela vida.

E eu não poderia deixar de agradecer o meu orientador, Prof. Dr. Pablo Faria, pela perseverança. Mesmo diante de tantos desafios, seja por eu ter vindo de outra área ou por eu ter dado tanta dor de cabeça, através dos problemas enfrentados nesse período árduo, se mostrou humano e me orientou, não só no que diz respeito às coisas da academia, mas em várias decisões importantes da vida, sempre paciente e de prontidão. Mais que um orientador, um grande amigo que ganhei durante esse percurso. Não teria como descrever sua importância, muito menos agradecer. Muito obrigado por tudo.

À Profa. Dra. Zenaide Carneiro, minha primeira orientadora. Agradeço não só pelas orientações, dicas sobre a vida, mas por me apresentar ao grande mundo da linguística. Agradeço o voto de confiança que me foi dado, ainda durante minha Iniciação Científica, ao me aceitar como orientando, no segundo semestre da graduação. Sem ela, esse caminho não existiria e eu ainda estaria "brincando" com os meus robôs no laboratório. O incômodo com a computação abriu uma porta, e novos caminhos puderam ser traçados a partir dela. Muito obrigado por tudo.

Agradeço também à minha mãe, Maristela Leal e ao meu irmão, Linoel Jr., que mesmo sem entender o que eu faço e todas as dificuldades que eu passei, sempre se mostraram prontos para me ajudar, não importando com o que fosse, sempre com "pode falar Igor, o que é?". E nessas conversas, os problemas se tornaram mais leves.

À minha tia Dilma, que acreditou que eu poderia chegar aonde estou, se mostrando confiante em mim, mais do que eu mesmo, e lembrando do que dizia meu grande tio Adauto Veiga "ele sabe que você consegue". Infelizmente, ele não pôde ver o que estou fazendo, mas sempre acreditou no "menino magrelo" e sua frase "ô Dilma, da comida para esse menino, ele tá com fome" me motivaram nessa caminhada.

À minha banca de qualificação, composta pela Profa. Dra. Ruth Lopes e pelo Prof. Dr. Marcelo Ferreira, que fizeram contribuições muito importantes, com críticas que me permitiram

melhorar a pesquisa e desenvolver um trabalho final. À Ruth, ainda agradeço, por suas aulas que pude assistir, sem dúvida, somaram para o desenvolvimento desta pesquisa.

Nas aulas da pós, me senti muito bem recebido pelos professores, pude experimentar um pouco mais do que é a linguística e tive a certeza de que estava no caminho certo. Não poderia deixar de agradecer a todas as professoras e todos os professores que estiveram presentes em minha jornada no IEL. Especialmente, à Profa. Dra, Charlotte Galves e à Profa. Dra. Lívia Oushiro, pela oportunidade de estágio na disciplina *Linguística de Corpus*, uma experiência singular e enriquecedora.

Fiz grandes amizades em Campinas, Rafa, Paulo, Francis, Caruzo, Stephanie, Amanda e Bea, que fizeram com que uma mudança de estado não tivesse nenhum peso, deixando sempre o clima mais leve e agradável. Ao Rafa, Paulo, Francis e Caruzo, especialmente, que foram as primeiras amizades feitas com o "churras" sempre movido a muita animação e descontração, fazendo com que nos sentíssemos em casa.

Agradeço também a todos os funcionários da Secretaria de Pós-Graduação do IEL, extremamente solícitos e pacientes, me auxiliando na resolução de problemas.

E com o fim dessa jornada, com as experiências vividas, espero começar outra em breve, o doutorado. Afinal, o que seria da vida sem as emoções (boas e ruins)? Vivemos para senti-las. Então, pego emprestado o título do livro escrito por Bilbo Bolseiro, ao retornar para o condado, *Lá e de volta outra vez*. É assim que me sinto, um pouco desgastado, mas pronto para começar tudo outra vez.

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

RESUMO

Neste trabalho, o objetivo principal é inserir uma componente morfológica em um modelo computacional de aprendizagem de categorias de palavras, baseado em informação distribucional, e avaliar seu impacto na performance do aprendiz computacional. Para tanto, o modelo utilizado para a inserção dessa componente linguística foi o desenvolvido por Faria e Ohashi (2018), que tem como inspiração o modelo apresentado em Redington *et al.* (1998). Ambos os modelos foram baseados na ideia da informatividade do contexto, vista em Harris (1954). Não é assumido, a priori, uma visão inatista ou empirista para o modelo, visto que esta tomada de posição não é relevante no contexto desta modelagem, que explora tão somente a informatividade dos dados de entrada, de tal modo que os resultados obtidos são úteis para ambas as perspectivas. Os resultados foram obtidos a partir de 48 condições experimentais, sendo aplicadas em 7 simulações distintas, nas quais são avaliadas diferentes decomposições morfológicas. As condições experimentais são divididas em 8 tipos e visam analisar aspectos diferentes que podem estar envolvidos no processo de aquisição: (i) janela de contexto; (ii) quantidade de palavras-alvo e de contexto; (iii) avaliação de performance por categoria; (iv) tamanho do corpus; (v) fronteira das sentenças; (vi) frequência e ocorrência; (vii) palavras funcionais; e (viii) o quanto uma categoria ajuda na categorização das outras. Os resultados foram analisados qualitativa e quantitativamente, além de serem avaliados quanto à sua significância estatística. Nosso resultados demostram melhora no desempenho do modelo quando introduzida a morfologia, se comparada ao modelo sem a morfologia. Além desse resultado, vale ressaltar que os resultados demostraram que a seleção das palavras-alvo impacta na categorização, que o uso da informação sobre a função do morfema não impacta no desempenho do modelo e que a informação morfológica sozinha se mostrou tão informativa quanto a informação com o contexto e morfologia.

Palavras-chave: modelagem computacional; aprendizagem distribucional; categorização de palavras; morfologia; aquisição da linguagem.

ABSTRACT

In this work, the main objective is to insert a morphological component in a computational model for learning word categories based on distributional information and to evaluate its impact on the computational learner's performance For that, the model used for the insertion of this linguistic component was the model developed by Faria and Ohashi (2018), which is inspired by the model presented in Redington et al. (1998). Both models are based on the idea of context informativeness found in Harris (1954). A priori, no innate or empirical view is assumed for the model, as this position is not relevant in the context of this modeling, which only examines the informativeness of the input data, so the results obtained are useful for both perspectives. The results come from 48 experimental conditions applied in 7 different simulations in which different morphological decompositions were evaluated. The experimental conditions are divided into 8 types and aim to analyze different aspects that may be involved in the acquisition process: (i) context window; (ii) set of target and context words; (iii) performance score by category; (iv) corpus size; (v) sentence boundary; (vi) frequency and occurrence; (vii) function words; (viii) and how much one category helps in categorizing the others. The results were analyzed qualitatively and quantitatively and additionally tested for statistical significance. Our results show an improvement in the performance of the model when the morphology is introduced compared to the model without the morphology. In addition to this result, it is worth noting that the results show that the choice of target words affects categorization, that the use of morpheme function information does not affect the performance of the model, and that morphological information alone proved to be as informative as information with context and morphology.
Key words: computational modeling; distributional learning; word categorization; morphology; language acquisition.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

Durante o processo de aquisição da linguagem, a criança tem à sua disposição pistas de diferentes naturezas que podem ser utilizadas para adquirir a língua, dentre as quais podemos citar as prosódicas, semânticas, fonológicas, sintáticas e morfológicas. No que diz respeito às pistas morfológicas, observadas neste trabalho, alguns estudos apontam que uma possível alavancagem sintática ocorre quando a informação distribucional do contexto é levada em consideração (FINCH; CHARTER, 1991, 1992). Aqui consideramos como *contexto* a vizinhança linear das palavras que serão categorizadas, por exemplo, em *A* ***gata*** *feliz*, o contexto da palavra *gata* na sentença seria *a* e *feliz*. Nessa linha, outros trabalhos demonstraram a utilidade da informação contextual para a categorização de palavras (REDINGTON *et al.* 1993; SHÜTZE, 1993; REDINGTON *et al.* 1995, 1998; MINTZ, 1995, 2002; FARIA; OHASHI, 2018). Os resultados apresentados por esses estudos demonstraram também que, embora a informação distribucional seja uma excelente pista para a categorização de palavras, entretanto, ela pode não ser suficiente para garantir o sucesso da criança nessa tarefa. Portanto, para contribuir com os estudos sobre o tema, novas pistas precisam ser consideradas.

Diante disso, com o intuito de auxiliar os estudos na área da modelagem computacional para a aquisição da linguagem, neste trabalho foram inseridas pistas morfológicas[1] ao modelo desenvolvido por Faria e Ohashi (2018). O objetivo geral é analisar o impacto da morfologia no modelo, com o intuito de responder a seguinte pergunta: o que acontecerá ao modelo quando inserirmos uma pista morfológica como variável linguística? Temos como hipótese básica que, ao inserir essa variável linguística, o desempenho do modelo seja melhor do que a sua versão anterior, uma vez que estamos inserindo uma variável que está disponível para o aprendiz ideal. E, como estamos utilizando um modelo computacional, é possível fazer experimentos que não seriam possíveis, ou muito difíceis, de aplicar nas crianças, conseguindo levar o aprendizado no modelo ao extremo. Então, como objetivos específicos, temos a avaliação do aprendiz computacional em nove experimentos, (1) variação do tamanho dos contextos; (2) variação do número de palavras alvo e de contexto; (3) avaliação do desempenho por categoria; (4) variação do tamanho do *corpus*; (5) marcação de fronteira das sentenças; (6) frequência *vs* ocorrência; (7) remoção das palavras

[1] Estamos considerando como *pistas morfológicas* a morfologia como variável no modelo. É levada em consideração uma simplificação do que observamos nos estudos morfológicos. Entendemos a complexidade da área, mas optamos por uma abordagem sóbria, principalmente, para conseguir controlar os efeitos dentro do modelo.

funcionais; (8) impacto de uma categoria sobre a outra na categorização; e (9) fala dirigida a criança *vs* fala entre adultos[2].

A presente dissertação está organizada em cinco capítulos. No **capítulo 1**, apresentamos os conceitos que direcionaram o desenvolvimento do trabalho, com uma breve explicação do que seria a modelagem computacional, seguida da apresentação do problema da categorização[3] de palavras e como o contexto pode auxiliar nessa tarefa, além de apresentar as diferenças entre a clusterização[4] que foi aplicada neste trabalho e os etiquetadores automáticos. Por fim, descrevo a visão geral da morfologia.

No **capítulo 2**, sintetizamos alguns trabalhos que utilizaram modelos similares ao que foi proposto aqui. Há modelos que utilizaram a perspectiva semântica e outros com a perspectiva sintática. Ambas as perspectivas são fundamentais na utilização da noção de contexto distribucional como pista, porém cada proposta tem suas particularidades, diante das diversas possibilidades de desenvolvimento que se abrem numa modelagem.

É no **capítulo 3** que apresentamos, detalhadamente, o modelo que foi utilizado como base e também como a morfologia foi acrescida ao modelo, adaptando-o para utilizar os dois tipos de informação (distribucional e morfológica). Também é explicamos o tratamento dado aos *corpora* e os parâmetros que foram utilizados para o desempenho do modelo.

No **capítulo 4**, discutimos os resultados obtidos, partindo dos resultados quantitativos das sete simulações realizadas, com o intuito de verificar o desempenho geral do modelo, comparando com as simulações trazidas por Faria e Ohashi (2018), Faria (2019a, 2019b). E, num segundo momento, apresentamos a análise qualitativa de três simulações, dando foco para aquelas que utilizam as informações morfológicas.

Finalmente, no **capítulo 5** apresentamos as *Considerações finais* deste trabalho, seguidas das referências bibliográficas.

---

[2] Esse experimento foi feito somente na simulação base, onde a morfologia não é levada em consideração.
[3] Categorizar no modelo significa clusterizar as palavras.
[4] Clusterizar, no modelo, significa agrupar as palavras seguindo os nossos critérios. Cf. capítulo 3, que trata do processo de clusterização.

1 MODELAGEM COMPUTACIONAL E CATEGORIAÇÃO

1.1 SOBRE A MODELAGEM COMPUTACIONAL

Antes de entrarmos na modelagem computacional apresentada neste trabalho, precisamos entender alguns aspectos da modelagem computacional. Podemos falar que uma das formas de conseguirmos alcançar resultados, dentro das mais diversas áreas de investigação científica, é simulando a realidade com o auxílio de computadores. A qualidade dessa simulação varia de acordo com a quantidade de informações que nós temos sobre o que será simulado e o quanto dessa informação podemos formalizar para inserir nos computadores, ou seja, quanto mais dados, melhor a simulação será.

Por exemplo, vamos imaginar a construção de um modelo que simula a queda de dois objetos. No primeiro momento, nós poderíamos observar o que está envolvido no processo, por exemplo, o peso de cada objeto e a força gravitacional, pois essa seria uma possível percepção inicial do movimento da queda. Após fazer isso, iríamos comparar o resultado da simulação com a realidade e observar as diferenças entre o modelo e o resultado real do experimento. E, em um segundo momento, aos poucos, mudaríamos a forma de analisar o problema da queda, inserindo novas variáveis e melhorando a simulação, até chegar ao ponto de estar de acordo, ou o mais próximo possível, da realidade. E, quando chegamos nesse nível de desenvolvimento, alcançamos um entendimento maior do problema, conseguindo uma resolução para o mesmo ou a manipulação e a avaliação do impacto das variáveis envolvidas durante o processo, de forma que esses dados sejam informativos para a realidade.

Ainda sobre modelos computacionais, Faria (2020) afirma que, para auxiliar na compreensão da realidade, um modelo precisa capturar a estrutura e a dinâmica essenciais do fenômeno, permitindo manipular e observar a interação de todas as variáveis potencialmente envolvidas no problema a ser investigado. Ou seja, não pode ser simplório a ponto de não conseguir evidenciar os principais pontos do que está sendo modelado, nem muito realista, a ponto de ser tão complexo quanto o fenômeno real.

No que diz respeito aos modelos utilizados para investigação da aquisição da linguagem, Pearl (2010) alerta que esses precisam ser desenvolvidos baseados em uma teoria específica e devem considerar as limitações da criança. Para isso, é preciso modelar levando em consideração níveis de processamento da informação, que inclui a descrição do problema a ser investigado (nível computacional), além de quais passos serão utilizados para solucionar o problema (nível algorítmico) e como o algoritmo será desenvolvido (nível implementacional). Ainda segundo a

autora, um modelo fornece uma maneira de investigar uma afirmação específica sobre a aquisição de linguagem, que envolverá uma questão informativa não óbvia. A potência da modelagem está na capacidade de especificar mecanismos de aquisição da linguagem hipotéticos sobre os quais temos total controle, de modo a permitir a manipulação das variáveis de interesse. Ainda, segundo Pearl (2010) (*apud* Marr 1982), os modelos devem atender a certos critérios avaliativos (suficiência formal, compatibilidade desenvolvimental e poder explicativo). Como explica a autora:

> A suficiência formal pergunta se o modelo aprende o que se espera dele, quando se espera dele, a partir dos dados disponíveis a ele. Isso é avaliado contra o que se conhece sobre o comportamento da criança e os dados de entrada. A compatibilidade desenvolvimental pergunta se o modelo aprende de uma forma psicologicamente plausível, usando recursos e algoritmos da mesma forma que uma criança faria. Isso é avaliado a partir do que se sabe sobre as capacidades cognitivas de uma criança. O poder explicativo pergunta qual é a parte crucial do modelo para que ele produza o comportamento correto e como isso impacta a afirmação teórica que o modelo está testando. Isso é avaliado pelo modelador por meio da manipulação das variáveis relevantes do modelo [...] (PEARL, 2010, p.166, tradução livre)[5].

Desse modo, quanto mais tais pré-requisitos forem satisfeitos, mais o modelo contribuirá significativamente para a pesquisa de aquisição da linguagem.

1.2 CATEGORIZAÇÃO DE PALAVRAS USANDO INFORMAÇÃO DISTRIBUCIONAL

A aquisição da sintaxe de uma língua envolve a aquisição de categorias sintáticas[6]. Como Redington (1998) alerta, explicar esse processo é uma tarefa difícil tanto para perspectivas inatistas quanto para empiristas. Do ponto de vista inatista estrito, teríamos um mapeamento das palavras-alvo para as categorias, já que estas seriam inatas. Porém, imaginamos que não seria produtivo um mecanismo de mapeamento que verificasse todas as combinações possíveis, uma vez que um mapeamento de 20 palavras para 2 categorias (2^{20}), sem restrição, resultaria em mais de um milhão de combinações. Como as línguas em geral possuem um número indefinido de palavras e muitas categorias mais, supomos que tal mapeamento seja baseado em algum tipo de regra interna (regras

[5] Formal sufficiency asks if the model learns what it is supposed to when it is supposed to from the data it is supposed to. This is evaluated against known child behavior and input. Developmental compatibility asks if the model learns in a psychologically plausible way, using resources and algorithms the way a child could. This is evaluated against what is known about a child's cognitive capabilities. Explanatory power asks what the crucial part of the model is for generating the correct behavior, and how that impacts the theoretical claim the model is testing. This is evaluated by the modeler via manipulation of the model's relevant variables (for example, whether the modeled children learn from unambiguous main-clause data only in the example above). (PEARL, 2010, p.166, trecho original)[5].

[6] A aquisição de categorias sintáticas no modelo pode ser entendida como o mapeamento de uma palavra para uma categoria.

gramaticais, por exemplo). Empiristas também precisam lidar com esse problema, uma vez que nesta perspectiva não há conhecimento prévio disponível para a criança. Portanto, além do mapeamento a criança precisa ainda descobrir as categorias sintáticas. Dessa forma, independente da visão assumida, descobrir as pistas que contribuem para a categorização sintática e que podem ser obtidas analisando os dados com o mínimo de assunções prévias, que é o caso da *informação distribucional*, nos permite começar a compreender como a criança tem sucesso nessa tarefa complexa.

Algumas críticas sobre a *informação distribucional* são levantadas por Pinker (1984). O autor afirma que: (i) a quantidade de relações distribucionais possíveis a considerar estaria fora do alcance de mecanismos de aprendizagem; (ii) muitas propriedades superficiais que um aprendiz leigo poderia explorar são irrelevantes, como mostram vários estudos linguísticos; (iii) mesmo dentre as propriedades relevantes, as línguas variam muito com relação a quais mobilizam; e (iv) correlações locais "espúrias" emergem em dados como "João come maçã" e "João come lentamente" (adaptação nossa), em que o aprendiz concluiria que "maçã" e "lentamente" são da mesma categoria. Entretanto, como apontam Redington *et al.* (1998), nenhum desses argumentos é convincente, pois: (i) não é preciso assumir que o aprendiz busca cegamente por qualquer propriedade possível; (ii) o fato de haver propriedades irrelevantes não impede que se aprenda com as que são relevantes; (iii) a variação entre línguas não pode ser obstáculo para este tipo de estudo, pelo contrário, o torna essencial; e (iv) cabe aos estudos mostrarem que o aprendiz pode superar os problemas locais, a partir de mecanismos psicologicamente plausíveis. Ademais, Redington *et al.* (*op. cit.*) ressaltam que esse é um problema tratável computacionalmente, o que torna viável sua investigação através de modelagem.

Somando aos argumentos acima, vale trazer aqui o estudo de caso de Valian (1986). Para investigar com quantos anos as crianças adquirem certas categorias sintáticas, a autora realizou um experimento com seis crianças, com idades entre 2;0 e 2;5, na tentativa de verificar se elas já teriam adquirido as seguintes categorias sintáticas: determinante, adjetivo, nome, sintagmas nominais, preposição e sintagmas preposicionais. Ainda de acordo com a autora, a importância de tal estudo é que, como as regras sintáticas são descritas em função das categorias, sem as categorias não teríamos a regras sintáticas. A partir de suas análises, Valian (*op. cit.*) conclui que com 2 anos e 5 meses, as crianças mostram um conhecimento vasto das categorias sintáticas, demostrando que a aquisição poderia ter ocorrido antes.

Valian (1986) aponta, ainda, que há algumas razões para acreditar que a aquisição dessas categorias sintáticas aconteceu por conhecimento sintático, não havendo correlação semântica óbvia. Além disso, o experimento mostrou que o aprendizado das categorias acontece de forma

gradual e que as crianças utilizam as regularidades distributivas como possíveis diagnósticos para a associação de uma categoria. As crianças procuram essas regularidades, concordâncias e outros fenômenos para testar as suas hipóteses. E, nesse caso, independe se é baseado em um conhecimento anterior de categorias ou não. Caso as crianças tenham um conhecimento prévio sobre categorias, utilizam da informação distribucional para alocar cada palavra em sua categoria, do contrário, elas utilizam essa pista como um guia que indica a existência dessas categorias e que precisam seguir essa regularidade. Isso indica, como argumenta Redington *et al.* (1998), que a investigação da aprendizagem distribucional pode escapar ao debate *inato versus não-inato*, sendo um mecanismo presente, qualquer que seja a assunção geral.

Desse modo, quando associamos uma determinada palavra a uma categoria sintática, utilizamos ao menos dois conhecimentos sem, necessariamente, perceber: a noção de que cada palavra pertence a uma categoria e a própria existência das categorias. Esses dois conhecimentos que colocamos em prática podem ser problemáticos para o aprendiz, pois mesmo que saibam da existência das categorias, ainda precisam identificar quais palavras pertencem a essas categorias. Então, uma possibilidade para auxiliar nessa dupla tarefa seria utilizar a informação do contexto das palavras, ou seja, aproveitar certas regularidades presentes na língua como pistas para essa categorização. O experimento de Valian (1986) mostrou essa tendência.

Tais regularidades no contexto das línguas foram estudadas por Harris (1954), ao analisar a coocorrência das palavras, propondo que a partir da análise dos contextos das palavras, é possível determinar um grau de similaridade entre elas. O autor defende que podemos supor uma relação de significado entre as palavras em três situações: (i) se duas palavras compartilham o mesmo contexto de forma quase idêntica, podemos concluir que elas são sinônimas; (ii) se compartilham parcialmente o contexto, a diferença de significado entre elas é a diferença entre seus contextos; e (iii) se as palavras não compartilham o contexto, são de categorias sintáticas diferentes. Desse modo, para direcionar a análise distribucional, Harris (1954) sugere considerar:

- Elemento: podemos dividir qualquer fluxo da fala em partes com tamanhos diferentes, variando de acordo com o que queremos encontrar. Como essas partes possuem uma certa distribuição, é possível encontrar as outras partes similares. E o que será encontrado varia de acordo com os interesses do pesquisador, sendo palavras, classes de palavras, morfemas ou fonemas;
- Similaridade: o quão próximos os elementos estão uns dos outros, sendo a similaridade definida de acordo com regras determinadas;

- Dependência: existe a possibilidade de que um conjunto de palavras A só ocorra quando em combinação com um conjunto de palavras B. Então, é plausível que tenhamos um elemento chamado AB;
- Substitutibilidade: dois elementos são substituíveis entre si se tem muita similaridade. E, de acordo com o grau de similaridade, podemos dividir entre sinônimo, diferente significado ou diferente classe;
- Domínios: são situações limitantes para que determinada estrutura apareça. Por exemplo, se uma palavra X acontece antes de uma sentença e isso implica que sempre ocorrerá determinada estrutura após esse aparecimento, temos a ideia de dependência; e
- Dados: as análises não podem ser feitas a partir de hábitos de falantes ou de algum gerador de padrões. Mas sim de um *corpus* real, onde se consiga derivar todas as regularidades que iremos estudar. Assim, podemos presumir que a regularidade estudada poderá ser encontrada em outros *corpora* semelhantes.

Tais discussões empíricas e teóricas são apoiadas por estudos computacionais como Finch e Chater (1991, 1992), Redington *et al.* (1993), Mintz (1995), Redington *et al.* (1995), Redington *et al.* (1998), Mintz e Newport (2002), entre outros. Cabe destacar que o modelo apresentado em Redington *et al.* (1998) foi utilizado como referência para o desenvolvimento do modelo apresentado aqui – no capítulo *3*, explicamos o modelo desenvolvido aqui, fazendo um paralelo entre os aspectos levantados por Harris (1954) e descrevendo como o modelo se relaciona com tais aspectos.

1.2.1 Etiquetar *versus* Clusterizar

Nesta seção, trazemos algumas diferenças entre *etiquetar* e *clusterizar*, já que são técnicas que podem ser utilizadas para o mesmo objetivo, mas aplicando procedimentos diferentes. Neste trabalho, utilizamos um *corpus* etiquetado morfologicamente – através do *eDictor*[7]– o *Corpus* Histórico do Português Tycho Brahe (CTB), como *corpus* de referência, já que o modelo apresentado aqui tem a *clusterização* como objetivo final e compara os grupos obtidos com as etiquetas empregadas no CTB.

[7] O eDictor (PAIXÃO DE SOUSA; KEPLER; FARIA, 2013) é um software desenvolvido para trabalhos filológicos e para análises linguísticas automáticas. O programa está disponível em: <https://humanidadesdigitais.org/edictor/>.

De modo geral, a *etiquetação* é um processo de anotação ao qual um texto é submetido, em que cada palavra representa um *token* que recebe uma *etiqueta* morfológica. Para facilitar esse processo, os etiquetadores podem ser treinados a partir de textos previamente anotados, podem operar com regras predefinidas ou, ainda, alguma outra metodologia que garanta um bom desempenho. As ferramentas de etiquetação disponíveis atualmente apresentam altos níveis de precisão e performance (em torno de 97% a 98% de acerto), oferecendo resultados satisfatórios do ponto de vista computacional, como pode ser visto em Christodoulopoulos *et al.* (2010), pois, ao final da aplicação, o importante é que o resultado se aproxime o máximo possível de 100% de precisão.

A *clusterização*, por outro lado, é um procedimento que não visa classificar e etiquetar palavras, mas sim reunir palavras que julga pertencentes à uma mesma classe em grupos (*clusters*), não havendo a relação palavra/etiqueta de forma direta como nos etiquetadores. Em outras palavras, enquanto a etiquetagem aplica etiquetas predefinidas à palavras, a clusterização apenas busca inferir agrupamentos possíveis para as palavras de um texto. Nos dois casos, a etiquetagem aplicada e os agrupamentos inferidos podem ter uma natureza mais semântica ou mais sintática, a depender do método e do objetivo.

1.3 A MORFOLOGIA

Como este trabalho trata da inclusão de informações morfológicas em um modelo computacional baseado na informação distribucional, nesta seção, tratamos da morfologia de forma não exaustiva, ou seja, sabendo que a discursão na área é ampla, não é o objetivo deste trabalho abordar tais questões. Sendo assim, no primeiro momento, trazemos uma perspectiva geral da morfologia a partir de Aronoff (2011) e, brevemente, uma perspectiva da morfologia no PB, a partir de Câmara Jr. (1970) e (MURATTI; MURATTI, 2011). Ou seja, o propósito aqui não é apresentar todas as nuances desse tema, mas trazer conceitos que foram utilizados nesta pesquisa para melhor entendimento do que vem adiante.

1.3.1 Visão geral da morfologia nas línguas humanas

Todos os seres humanos são capazes de entender e criar novas palavras, seja uma palavra completamente nova ou uma nova palavra utilizando afixos, unindo-os com palavras já existentes. Esse fenômeno pode ser explicado pela existência da morfologia, que do ponto de vista linguístico,

segundo Aronoff (2011), pode ser definida como um sistema mental envolvido na formação de palavras ou como o estudo da estrutura da formação dessas palavras, a partir do qual podemos fazer dois tipos de análises, a analítica e a sintética.

A *analítica* é uma abordagem na qual todas as palavras são analisadas, desmembradas e comparadas entre si, a partir de uma amostra de dados. De forma geral, esse tipo de análise sempre é importante a priori, por ser na qual observamos os dados sem uma perspectiva teórica. Após desmembrar as palavras, é necessário entender cada uma das suas partes, como funcionam e como podem ser formadas, processo que chamamos de *síntese*. Para ilustrar esse processo, tomemos o exemplo trazido por Aronoff (2011), que faz a comparação com a desmontagem e remontagem de um relógio: na desmontagem, serão encontradas diversas peças que são necessárias para a remontagem; e se essa remontagem for feita por tentativa e erro, talvez não seja possível montar novamente o relógio ou demore muito mais tempo. Então, seria mais fácil ter uma espécie de teoria que guie esse processo. Fazendo um paralelo com a morfologia, o fato de conseguirmos criar novas palavras significa que temos um guia, que mostra o que está envolvido no processo de manipulação dos morfemas. Dessa forma, os seres humanos usariam as duas análises, a *analítica*, quando se deparam com uma nova palavra, separando para entender o que está ali. E a *síntese*, quando formam novas palavras, ou seja, entendendo o processo de criação.

Quando essas análises são realizadas, dois processos podem estar envolvidos, a *flexão* e a *derivação*, ambos com a utilização de *afixos* no processo de modificação das palavras, sendo esses afixos, morfemas que podem ser unidos a outros radicais, no início ou final da palavra (*prefixos* para o início e *sufixos* para o final[8]). No processo de *flexão*, há modificação nas formas gramaticais das palavras, como em *ox/boi, ox-en/bois* (singular, plural), como exemplificado em Aronoff (2011, p.47). Basicamente, a morfologia flexional adapta palavras existentes a seu contexto sintático e a aspectos semânticos, mas sempre mantendo o sentido básico da raiz. No processo de *derivação*, ocorre a criação de palavras a partir de outras, com mudança do significado em maior ou menor grau e até mesmo da categoria da palavra inicial. Por exemplo, temos *tie/amarrar* e *un-tie/desamarrar*, em que acontece a criação de um antônimo, mantendo a classe verbal. Já em *digg/cavar* e *digg-er/escavador*, vemos a criação de um *substantivo* a partir de um *verbo*. Quando há as realizações fonológicas dos morfema, denominamos morfe. Por exemplo, para o inglês, Aronoff (2011) traz os exemplos do morfema *-ed*, que pode ter o morfe *t* em *jumped* (pulou) e pode o morfe *d* em *repelled* (repeliu).

[8] Há outros tipos de afixos atestados nas línguas, como infixos e circunfixos, que optamos por não discutir aqui por não serem diretamente relevantes para o presente estudo.

1.3.2 Visão geral da morfologia no português brasileiro

No que diz respeito ao PB, é possível observar a ocorrência de *morfema zero*, *morfes cumulativos*, *morfes alternantes*, *morfes redundantes* e *morfes homônimos* (MURATTI; MURATTI, 2011). O *morfema zero*, representado pelo símbolo Ø, acontece quando existe um morfema que marca gênero ou plural, por exemplo, e não existe a mesma marca para o oposto, como em professor-a / professor-Ø. Os *morfes cumulativos* são aqueles que carregam mais de uma função, como no caso de cantá**-sse-**mos, onde -sse- indica o *tempo* imperfeito e *modo* subjuntivo. Os *morfes alternantes* são aqueles que ocorrem na troca entre dois fones, como em av**ô**/av**ó**, e são de natureza vocálica, consonantal ou suprassegmental, como em f**i**ro/f**e**res, di**g**o/di**z**es e ret**í**fica/ret**i**fica. Quando há um morfema que marca o plural ou o gênero, por exemplo, e ainda assim existe uma alternância na raiz, são chamados de *morfes redundantes*, como em p**o**ço/p**o**ços, s**o**gro/s**o**gra e tra**g**o/tra**z**es. Por fim, os *morfes homônimos*, que são morfes onde há a mesma forma, mas significado diferente, como em (o) cant**-o**/(eu) cant**-o**, sendo a primeira palavra nome e a segunda verbo.

Como dito anteriormente, é possível, a partir do processo de derivação, criar novas palavras utilizando os morfemas derivacionais. Os morfemas derivacionais que ocupam a posição anterior ao radical em PB, os prefixos, têm algumas características, como a mudança do significado da raiz, em algumas situações, como em ***in-***certeza, não indicando categorias gramaticais de gênero, número, tempo, modo e pessoa; também não alteram as categorias sintáticas, os nomes continuam sendo nomes, os verbos continuam sendo verbos, como em fazer/**re-**fazer. Ainda, alguns prefixos ganham autonomia morfológica, como em **extra-**ordinários, por exemplo. Já os sufixos apresentam características diferentes, não alteram a significação da raiz, como em *sapat**-o**/sapat**-eiro***, e, ao contrário do prefixo, podem alterar as categorias sintáticas, como em *real/real**-izar***.

Os morfemas flexionais ou desinências são aqueles que mudam a flexão dos verbos, nomes, adjetivos, pronomes, artigos e numerais. As desinências no PB são classificadas em *nominais*, que indicam gênero (masculino e feminino) e número (singular e plural), e verbais, classificadas em modo-temporal (tempo e modo do verbo) e número-pessoal (número e pessoa da forma verbal).

Sobre as desinências nominais, Câmara Jr. (1970) aponta que a descrição de gênero dos nomes pode ser feita como *nomes de gênero único* ((a) rosa); *nomes de dois gêneros sem flexão*(o, a) artista); e *nomes de dois gêneros*((o) lobo/(a) loba). Além disso, os nomes podem, ainda, ser classificados considerando a *heteronímia*, quando há representação de gênero usando palavras distintas (pai/mãe), a *derivação sufixal*(galo/galinha) e a *alomorfia no radical* (frade/freira). Quanto à flexão de número, esta diferencia o singular e o plural, como dito. Entretanto, em alguns casos, o plural marcado nos

nomes (-s) não indica mais de uma ocorrência, sendo apenas uma marcação gramatical, sem oposição de sentido, como em *os óculos*; além de alguns nomes que mudam o sentido quando flexionados (a honra/as honras). Desse modo, nem sempre podemos afirmar que a presença do *-s* indica necessariamente o plural, sendo que existem situações em que há modificações morfofonêmicas, como em limão/limões.

Os verbos têm uma complexidade flexional maior que a dos nomes, dada a quantidade de tempos, modos e pessoas verbais – há pelo menos 13 tempos verbais, além de 3 pessoas para o singular e 3 pessoas para o plural. Uma possível representação para os verbos é *radical* + *vogal temática* + *desinência modo temporal* + *desinência número pessoal*, mas não significa que todos os verbos apresentarão todas os morfemas. Por exemplo, *trocam*, que não tem a desinência modo temporal, sendo marcada com o símbolo Ø, indicando vazio: troc-(radical) ,-a- (vogal temática), -Ø (desinência modo temporal) e -m (desinência número pessoal). Além disso, ainda existem os verbos irregulares, que não seguem o padrão geral dos verbos, como, por exemplo, o verbo *ir*, que na primeira pessoa do presente do indicativo é *vou*, tendo o radical *v-* e a desinência número pessoal *-ou*, os demais afixos são marcados com o -Ø.

Dito isso, podemos perceber que a representação morfológica pode acontecer em vários níveis, por exemplo, podemos falar apenas dos afixos, sem falar necessariamente das suas funções morfológicas, ou, ainda, lidar apenas com os prefixos ou sufixos sem uma decomposição exaustiva. Essas possibilidades, por parte da morfologia, nos colocam à frente de várias decisões quando estamos lidando com modelos computacionais, pois, precisamos definir qual o nível da análise vamos inserir no modelo. Afinal, mesmo sabendo que é possível um nível de análise mais detalhado, nem sempre é o que o modelo precisa. No capítulo 3, onde discutimos o modelo aqui utilizado, explicamos as decisões tomadas para a escolha do nível da morfologia que foi utilizado.

2 ESTUDOS SOBRE O TEMA

Conforme apresentado anteriormente, a modelagem computacional nos auxilia no desenvolvimento e na compreensão dos fenômenos envolvidos em um determinado processo. Nesta seção, apresentamos estudos que têm como parte da sua metodologia o uso de modelos computacionais, que analisam o contexto das palavras, com o intuito de encontrar relações linguísticas e, alguns, com o intuito de analisar como tais contextos podem ser uma possível pista para o entendimento do processo de categorização ou ainda, quais informações podemos retirar a partir da regularidade encontrada entre eles.

Cabe apontar que, na investigação científica, podemos nos deparar com estudos que, ainda que possuam objetos de pesquisas semelhantes, trazem diferentes objetivos e perspectivas de análise. Na modelagem computacional, não é diferente. Então, mesmo que esta pesquisa tenha como objetivo a análise do contexto das palavras, com foco na categorização de palavras, nós nos deparamos com estudos que se diferenciam tanto em aspectos técnicos quanto metodológicos. No leque dos modelos que analisam a informação distribucional de contexto, irei apresentar dois grandes grupos, os que trabalham com a perspectiva semântica e os que trabalham com uma perspectiva sintática. Nesses grupos, há modelos que não utilizam as informações com a intenção de investigar o impacto linguí, apenas analisam a relação das palavras em diferentes textos, no geral, com o intuito de melhorar métodos de processamentos de linguagem natural.

2.1 OS MODELOS FOCADOS NA SEMÂNTICA

Entre os estudos que trabalham com a perspectiva semântica, Deerwester *et al.* (1990) trazem uma proposta de investigação que utiliza o contexto das palavras para criar um método de recuperação de informações baseado na semântica. O foco desse método é melhorar a indexação na busca de documentos, substituindo a busca que eles chamam de *tradicional*, onde o retorno do documento acontece a partir da combinação entre duas palavras iguais, por um método que consegue retornar, por exemplo, um documento mesmo que o termo pesquisado não esteja explicitamente nele, utilizando a análise semântica latente (LSA). O modelo proposto por Deerwester *et al.* (1990) é baseado em uma tabela com a relação *Termos x Documentos*, onde cada termo é representado por um vetor; e para encontrar a relação entre os documentos, aplicam a decomposição de valores singulares (SVD). A partir desse resultado, é possível criar uma relação entre um termo a ser pesquisado e o documento. Por exemplo, ao procurar por *barco*, mesmo que

um documento não apresente tal palavra explicitamente, mas sim termos como *proa* (remetendo a barco), tal documento poderia ser retornado no resultado da busca. Os resultados obtidos se mostraram superiores em um teste (precisão de 0,51 *versus* 0,45) e igual em outro, se comparados ao modelo *tradicional*. Segundo os autores, é o suficiente para um encorajamento para o novo método, já que é possível melhorar fazendo um tratamento a priori no texto e utilizando um sistema de busca voltado para ele.

Já Landauer e Dumais (1997) investigaram, também utilizando os métodos LSA e SVD, se era possível um modelo julgar novas palavras como sinônimas de outras, para isso, aplicaram esses métodos em um *corpus geral*[9] e tomaram, como base para avaliação do modelo, os resultados obtidos no Teste de Inglês como Língua Estrangeira (TOEFL) realizado por pessoas. Além disso, verificaram a taxa de aprendizado de novas palavras, comparando o resultado do método com o de crianças em idade escolar. O teste do TOEFL era composto por 80 julgamentos de múltipla escolha entre uma palavra-alvo e quatro significados. Nesse modelo, os pesquisadores utilizaram o contexto das palavras para formar uma matriz composta por *Palavras-alvo x Contexto*. Porém, como é alertado por Landauer *et al.* (1998), nesse método o contexto não trata de uma análise baseada em uma contagem simples do contexto das palavras, na qual apenas as palavras são unidades, mas sim de um contexto complexo, que envolve até sentenças como unidades de contexto. A partir dessa matriz gerada, o modelo utiliza a técnica SVD para encontrar uma determinada medida, o fator do SVD, que pode ser utilizada como métrica para indicar se duas palavras estão próximas ou não. Quanto maior o fator SVD, maiores as chances de essas palavras serem sinônimas. O modelo se mostrou capaz de resolver os testes de significados do TOEFL, apresentou uma taxa de 64,4% de palavras corretas *versus* 64,5% de palavras incorretas, e encontrou uma taxa de aprendizado de novas palavras próximo ao das crianças em idade escolar.

Uma abordagem diferente das apresentadas até o momento pode ser vista em Bullinaria e Levy (2007), que fazem quatro experimentos para avaliar o quanto de informação semântica pode ser extraída da coocorrência dos contextos: (i) sinônimos utilizando o teste do TOEFL; (ii) análise da medida de distância entre as palavras; (iii) capacidade de análise da distância semântica entre as palavras; (iv) e se há possibilidade de utilizar a mesma análise para verificar as categorias sintáticas. Nessa proposta, os autores utilizaram uma contagem de contexto simples, na qual a janela de contexto, que é a quantidade de palavras que serão analisadas ao redor da palavra-alvo, é variável. Tal janela é utilizada para a construção de um vetor com as ocorrências dos contextos para uma determinada palavra-alvo. Expandindo o que foi feito em Bullinaria e Levy (2007), Bullinaria e Levy (2012) continuaram com a mesma abordagem, porém aumentando a quantidade de experimentos,

[9] Utilizamos o termo *corpus geral* para fazer referência ao *corpus* que não é de fala dirigia à criança.

inserindo a análise do SVD, verificando a informação das palavras funcionais, entre outros experimentos. Como o intuito era de avaliar diferentes métricas sobre métodos diferentes de extração semântica de coocorrência que nos permite verificar a performance das métricas utilizadas. Por exemplo, no mesmo experimento do TOEFL, feito por Landauer e Dumais (1997), em uma das variações apresentadas no artigo, foi obtida uma porcentagem de acerto de 72,5% *versus* os 64,4% de Landauer e Dumais (1997).

Também a fim de demonstrar a utilidade do contexto das palavras, mas aplicado em redes neurais voltadas para semântica, Kajic e Eliasmith (2018) inseriram essa propriedade para demonstrar que tais redes que a utilizam são tão capazes quanto outras, que utilizam metodologias distintas. Para isso, os estudiosos utilizaram dois modelos de redes neurais, a Word2Vec e a GloVe, empregados para processamento de linguagem natural. Em comparação com as demais, ambas as redes também se saíram bem, assim como as demais, conseguindo construir o que eles chamaram de *conhecimento de pequeno mundo*, ou seja, conseguiram encontrar o que está sendo buscado, navegando por poucos nós da rede.

Há, também, modelos que tentam encontrar na análise distribucional padrões semânticos ao redor de determinadas palavras, como é o caso de Pado e Hole (2019), que utilizaram esse tipo de modelo para analisar a polissemia no pronome alemão *sich*, se mostrando útil, segundo os autores, uma vez que conseguiram observar e analisar o pronome nas 8 posições previstas, a partir do resultado obtido pela rede neural. Dessa forma, tanto esse modelo quanto os outros demonstrados, até o momento, se mostraram eficazes em solucionar determinados problemas, utilizando apenas o contexto das palavras, evidenciando que existe um conhecimento a ser explorado nesse tipo de informação e a possibilidade para uma *alavancagem semântica*.

2.2 OS MODELOS FOCADOS NA SINTAXE

Há modelos como o de Schutze (1993), que utilizam uma rede neural para categorizar as palavras. Nesse tipo de modelo, não temos o controle do que acontece entre os neurônios da rede, porém, podemos configurá-los de forma que o aprendizado aconteça, baseado nas regras definidas. Além disso, as regras aplicadas nessa rede se baseiam no *corpus* utilizado e, também, utilizam apenas o contexto das palavras como pistas. Os resultados obtidos pela rede neural se mostraram efetivos no teste feito pelo autor.

Não podemos deixar de falar dos precursores do modelo apresentado em Redington *et al.* (1998)[10] e Finch e Chater (1991, 1992), que trazem a proposta de verificar se é possível categorizar palavras sem conhecimentos sintáticos prévios, apenas utilizando o contexto de coocorrência entre as palavras. Nesses estudos, foi trazida a possibilidade de o contexto ser utilizado como uma pista que ajudaria na *alavancagem sintática*. Para essa verificação, os autores propuseram um sistema híbrido, no qual a similaridade entre as palavras é calculada por uma rede neural, e os resultados obtidos são utilizados para que um modelo simbólico seja o responsável pelo agrupamento das palavras, nesse caso, agrupando as palavras com as medidas de similaridade mais próximas em um dendrograma. Nesse tipo de representação, é possível observar níveis de agrupamentos diferentes, nos quais as palavras com maior proximidade são consideradas pertencentes à mesma categoria e as palavras mais distantes são consideradas de categorias diferentes. Os autores não utilizaram uma medida para a avaliação do modelo, porém apresentaram como resultado alguns agrupamentos obtidos, demostrando a efetividade do método proposto.

Dando continuidade aos modelos vistos em Finch e Chater (1991, 1992), temos a proposta desenvolvida por Redington *et al.* (1993). Nesse novo modelo, os pesquisadores não implementaram uma rede para encontrar a medida de similaridade, mas fizeram um cálculo utilizando os vetores que armazenaram os contextos das palavras-alvo, denominado de *vetor de contexto*. Além da mudança no modo como a similaridade foi calculada, nesse estudo utilizaram como entrada o *corpus* CHILDES (MACWHINNEY, 1989) a fim de aproximar os dados de entrada aos dados dos aprendizes. O estudo visou, ainda, categorizar as palavras sem o auxílio de outras variáveis além do contexto, alcançando grupos com 90% de palavras categorizadas corretamente. Cabe destacar que outra variação desse estudo pode ser vista no modelo desenvolvido em Redington *et al.* (1995), quando os pesquisadores utilizaram a mesma abordagem com um *corpus* de fala chinês, alcançando uma certeza de categorização das palavras nos grupos encontrados de até 70%.

Outro modelo que utiliza métodos semelhantes aos encontrados em Finch e Chater (1991, 1992) é o modelo proposto por Mintz *et al.* (1995), que traz como proposta avaliar a informatividade dos contextos das palavras, categorizando verbos e nomes, utilizando apenas as palavras anterior e seguinte da palavra-alvo a ser categorizada. Os autores apontaram que uma das justificativas para essa investigação seria o fato de que, mesmo que as crianças tenham conhecimento inato das classes, elas não saberiam de antemão quais palavras da língua dela pertencem a determinada categoria. Ademais, a *informação distribucional* pode ser uma pista que está sempre à disposição nos

[10] O modelo apresentado por Redington *et al.* (1998) foi utilizado como inspiração do modelo proposto neste trabalho.

dados de entrada. E mesmo que a informação distribucional por si não seja suficiente para a aquisição das categorias, ela pode ser uma fonte de informações úteis, como nos resultados demonstrados no estudo.

Dando sequência ao modelo, Mintz *et al.* (2002) expandiram os experimentos com o propósito de tornar o *input* do modelo mais parecido com o *input* do aprendiz. Para isso, inseriram fronteiras nas sentenças, substituindo classes fechadas por um símbolo (FNCT) que sinaliza a presença daquela palavra na sentença; reduzem o tamanho do *input*, diminuindo a quantidade das palavras que podem ser utilizadas como pistas na janela de contexto; e testam novas janelas de contexto para verificar se possíveis ruídos prejudicam o modelo. Mesmo com essas variações, foi possível obter um resultado positivo, principalmente pelo objetivo do método, que é analisar a informação distribucional do *input* e não, necessariamente, modelar perfeitamente uma criança. Nos resultados obtidos por Mintz *et al.* (2002), foi possível encontrar agrupamentos com uma pureza de até 80%, sendo esse percentual o indicativo da quantidade de palavras de apenas uma categoria sintática no mesmo grupo.

Em Clark (2003) podemos ver um modelo que insere a morfologia como outra pista para a categorização de palavras utilizando a análise distribucional em 7 línguas diferentes (inglês, búlgaro, tcheco, estônio, húngaro, romeno e esloveno). Para esse experimento, o autor utilizou, além da clusterização, experimentos que levam em consideração a frequência e um algoritmo bayesiano para agrupamento de palavras que são tidas como pertencentes ao mesmo grupo. Nos outros modelos apresentados até aqui, tínhamos variações, mas todos utilizaram apenas o contexto da palavra-alvo para que fosse possível a categorização. Em um dos experimentos que Clark (2003) propos, uma palavra só pode fazer parte do *cluster* se ela tiver algum padrão morfológico com as outras palavras envolvidas, como, por exemplo, palavras terminadas em *-ing*. Para fazer essa análise, ele criou uma fórmula que calcula a probabilidade para que um subconjunto de caracteres pertença ao mesmo grupo, dessa forma, analisando a morfologia. Comparando com outros experimentos feitos por ele, a morfologia se mostrou um diferencial nas palavras com frequências menores que 5 no *corpus*, para 5 das 7 línguas utilizadas; e para as outras 2, a morfologia com a frequência se mostrou melhor.

Por fim, um estudo que se aprofundou mais na questão morfológica, dentre os modelos que utilizaram a informação distribucional, foi o de Onnis e Christiansen (2008). Os autores desenvolveram um modelo sensível aos padrões nas bordas das palavras, com a intenção de categorizar as palavras em nomes, verbos e uma categoria outras, onde colocaram as palavras que não são verbos ou nomes. No total, foram feitos 6 experimentos, dentre os quais, 3 tiveram o objetivo de avaliar mais detalhadamente o comportamento no inglês e 3 foram destinados a outras

línguas (alemão, francês e japonês), utilizando a os dados do CHILDES. Como base para esses experimentos, foi apresentado um contexto diferente, no qual o que é contabilizado não são as palavras ao redor de uma determinada palavra-alvo, mas sim os afixos dela, sendo que cada afixo é representado por uma coluna na tabela das palavras que seriam categorizadas. Para o experimento 1, as palavras foram decompostas utilizando o CELEX[11] como base para a extração dos afixos. Nesse experimento, foi possível categorizar corretamente 60,7% no geral, sendo o melhor desempenho na categorização dos verbos. O experimento 2, baseado nas bordas fonológicas das palavras, feito utilizando o CELEX; as palavras foram transcritas foneticamente e o menor segmento[12] foi extraído das bordas, por exemplo, *does* (/dʌz/), onde foram contabilizados /d/- e -/z/ como bordas da palavra. O resultado geral foi de 59,7% de categorização correta, também tendo uma vantagem para os verbos na categorização. O experimento 3 foi baseado em uma rede neural não supervisionada a partir da decomposição utilizada no experimento 2. Como resultado, obtiveram uma categorização geral correta acima dos 45% e melhor categorização para o grupo *outras*. Os experimentos 4, 5 e 6 seguiram os mesmos procedimentos do experimento 2, sendo que o experimento para o alemão alcançou 54% de acerto, com melhor categorização para os verbos. O experimento para o francês obteve 53,9% de acerto geral, com os verbos mais bem classificados. E o experimento para o japonês alcançou uma categorização correta 51,5% de forma geral, com os verbos com melhor taxa de categorização. Assim, observamos que todos os experimentos obtiveram uma porcentagem de categorização maior que a taxa base, que alcançou, no máximo, 33% no experimento 2. Em todos os experimentos demonstrados até aqui, a morfologia se mostrou uma pista com resultados melhores do que o resultado de base, mas não o suficiente para alcançar uma categorização perfeita em conjunto com a *informação distribucional*.

[11] Banco de dados com informações ortográficas, fonológicas, morfológicas, sintáticas e frequência de palavras: < https://catalog.ldc.upenn.edu/LDC96L14>.

[12] Optamos por utilizar a expressão *menor seguimento* diante da escolha dos autores em "By selecting the smallest phonological unit [...]" (Onnis e Christiansen, 2008, p. 195), para indicar que estava extraindo seguimentos da borda.

3 O MODELO DISTRIBUCIONAL

Conforme já pontuado ao longo do texto, os problemas precisam seguir uma série de passos para serem modelados computacionalmente e, além das possíveis limitações impostas por esses passos, precisam ter uma representação da realidade de forma que possamos extrair algum tipo de informação. E para a aquisição, ainda precisam, minimamente, atender aos critérios de Pearl (2010), listados na introdução deste trabalho. Sendo assim, nesta seção, demonstramos como foi feita a modelagem da morfologia, o modo como foi inserida no modelo, além de explicar o funcionamento do modelo sem a morfologia inserida, descrevendo os modelos que tomamos como base, com a intenção de deixar o mais simples possível o entendimento do que está sendo proposto.

O modelo simulou uma aquisição instantânea, baseado em um aprendiz que armazena informações e, após uma certa quantidade de dados armazenados, analisou e tomou decisões seguindo alguns critérios, foram eles: palavra a ser categorizada (palavra-alvo), tamanho da vizinhança da palavra alvo (janela de contexto) e quais palavras ao redor da palavra alvo serão analisadas (palavras de contexto). O modelo base não utilizou a morfologia como parte da tomada de decisões, mas utilizou como palavras-alvo as 1000 palavras mais frequentes do *corpus*, sendo as 150 mais frequentes utilizadas como palavras de contexto; e a janela de contexto foi definida como as duas palavras anteriores e as duas palavras seguintes. Quanto ao modelo com a morfologia, seguimos a mesma configuração para as palavras de contexto, mas utilizando somente 650 palavras-alvo, isso aconteceu porque nem todas as 1000 palavras mais frequentes passaram no teste para receber informação morfológica. Essas informações e outras a respeito do modelo serão descritas a seguir.

3.1 A BASE DO MODELO

Para o modelo apresentado aqui, conforme já citado, utilizaremos o modelo desenvolvido em Faria e Ohashi (2018), Faria (2019a) e Faria (2019b) e tendo como estudos base, Redington *et al.* (1998), Mintz (2002), Clark (2003) e Onnis e Christiansen (2008). Dessa forma, com o intuito de demonstrar que as propriedades distribucionais das palavras podem ser altamente informativas, no que diz respeito à categoria sintática, e como essa informação pode ser extraída por alguns mecanismos psicologicamente plausíveis, Redington *et al.* (1998) propôs três estágios para desenvolver esse tipo de análise:

(i) medir os contextos de distribuição em que cada palavra ocorre;

(ii) comparar o contexto de distribuição para pares de palavras; e

(iii) agrupar palavras com distribuições de contextos similares.

O primeiro estágio envolve coletar o contexto (por exemplo, duas palavras imediatamente seguintes e anteriores a palavra-alvo, *Tabela 1*) em que as palavras ocorrem, isto é, são coletadas estatísticas de coocorrência entre a palavra-alvo e as palavras em seu entorno, armazenando os dados estatísticos em uma *tabela de contingência* (*Tabela 2*), na qual cada célula registra a frequência de coocorrência da palavra alvo com uma dada palavra de contexto. Uma representação gráfica pode ser vista na *Figura 1*. Após a coleta desses dados, cada linha da tabela de contingência é uma linha correspondente a uma palavra alvo, formando uma representação vetorial da distribuição observada das palavras de contexto na posição relevante, chamado de *vetor de contexto* (*Tabela 3*).

Tabela 1: Exemplo de uma janela de contexto em que são contabilizadas duas palavras anteriores da palavra alvo e duas depois. Nesse exemplo, temos *uma* como a palavra alvo. O termo *palavra de contexto* foi abreviado para PC e ao lado da abreviatura a posição que ela representa na tabela de contingência. Nesse caso, ficou em ordem crescente apenas para facilitar a exemplificação.

Janela de contexto							
-	Alvo -2 (PC_1)	Alvo -1 (PC_2)	Palavra alvo	Alvo +1 (PC_3)	Alvo +2 (PC_4)	-	-
A	Belldandy	é	uma	gata	bonita	e	teimosa

Tabela 2: Representação de uma tabela de contexto, onde as linhas são os vetores de contexto que formam as palavras-alvo (PA).

Tabela de contingência				
Palavra alvo	Posição na janela de contexto	Ocorrências PC_1	...	Ocorrências PC_n
PA_1	Alvo -2	0...n	0...n	0...n
	Alvo -1	0...n	0...n	0...n
	Alvo +1	0...n	0...n	0...n
	Alvo +2	0...n	0...n	0...n
PA_2	Alvo -2	0...n	0...n	0...n
	Alvo -1	0...n	0...n	0...n
	Alvo +1	0...n	0...n	0...n
	Alvo +2	0...n	0...n	0...n
...	...	...	...	...
PA_n	Alvo -2	0...n	0...n	0...n
	Alvo -1	0...n	0...n	0...n
	Alvo +1	0...n	0...n	0...n
	Alvo +2	0...n	0...n	0...n

Tabela 3: Representação final da palavra-alvo em um vetor de contexto.

Vetor de contexto				
PA_n	Vetor Alvo -2	Vetor Alvo -1	Vetor Alvo +1	Vetor Alvo +2

O segundo estágio envolve avaliar a similaridade entre os *vetores de contexto*. Segundo Redington et al. (1998), o vetor de contexto geral para cada palavra alvo pode ser pensado como um ponto em um espaço de possíveis distribuições de palavras (em função do contexto). Assim, é

possível esperar que as palavras com a mesma categoria sintática tenham distribuições similares. Para medir essa similaridade, os autores utilizaram o *coeficiente de correlação de postos de Spearman (ρ)*[13], cujos valores variam entre X e Y, sendo que quanto mais alto, maior a similaridade, sendo que os valores variam entre -1 e 1, indicando correlação negativa ou positiva. No modelo, os valores obtidos são posteriormente normalizados para o intervalo de 0 a 1, sendo que quanto mais alto, maior a similaridade no modelo.

O agrupamento a partir da similaridade está relacionado com o terceiro estágio. Para isso, os autores utilizam a *análise de cluster hierárquica padrão* (SOKAL; SNEATH, 1963, apud REDINGTON et al., 1998), conhecida como *cluster de link médio*. O algoritmo começa combinando itens que estão mais próximos, no nosso caso, palavras-alvo, de acordo com a métrica de similaridade. Uma vez que os itens são combinados, um *cluster* é formado, podendo ser ele mesmo agrupado em conjunto com palavras-alvo próximas ou com outros *clusters*. A distância entre dois *clusters* é a média das distâncias entre os membros de cada um. Ao final, teremos um *cluster* que representa uma estrutura hierárquica entre os vários *clusters* encontrados. Assim, a quantidade de grupos encontrada vai depender do local em que o dendrograma é cortado, em outras palavras, qual ponto de similaridade do dendrograma iremos selecionar. O agrupamento hierárquico final pode ser visto na *Figura 1*.

A fim de verificar como a análise distribucional se comporta diante das diversas condições de aquisição da linguagem da criança, Redington *et al.* (1998) simulou nove diferentes experimentos: (i) variou o contexto; (ii) variou a quantidade de palavras-alvo, (iii) avaliou a precisão do método para cada classe de palavras; (iv) variou o tamanho do *corpus*; (v) avaliou o método utilizando diferentes formas de marcar fronteiras de enunciado; (vi) avaliou a eficácia do método substituindo a frequência das palavras de contexto pela informação binária de sua ocorrência; (vii) avaliou o método retirando as palavras funcionais do *corpus*; (viii) avaliou se a categoria de uma palavra de contexto é mais informativa que a palavra em si; e (ix) avaliou se há diferença na análise distribucional, quando se contrasta a fala dirigida à criança com a fala entre adultos.

Em Mintz et al. (2002), experimentos similares foram desenvolvidos. Inicialmente, os autores analisaram o contexto de uma palavra a partir da palavra imediatamente anterior e seguinte. Nesse método, o tipo similar de *tabela de contingência* utilizada em Redington *et al.* (1998) é construído. Para

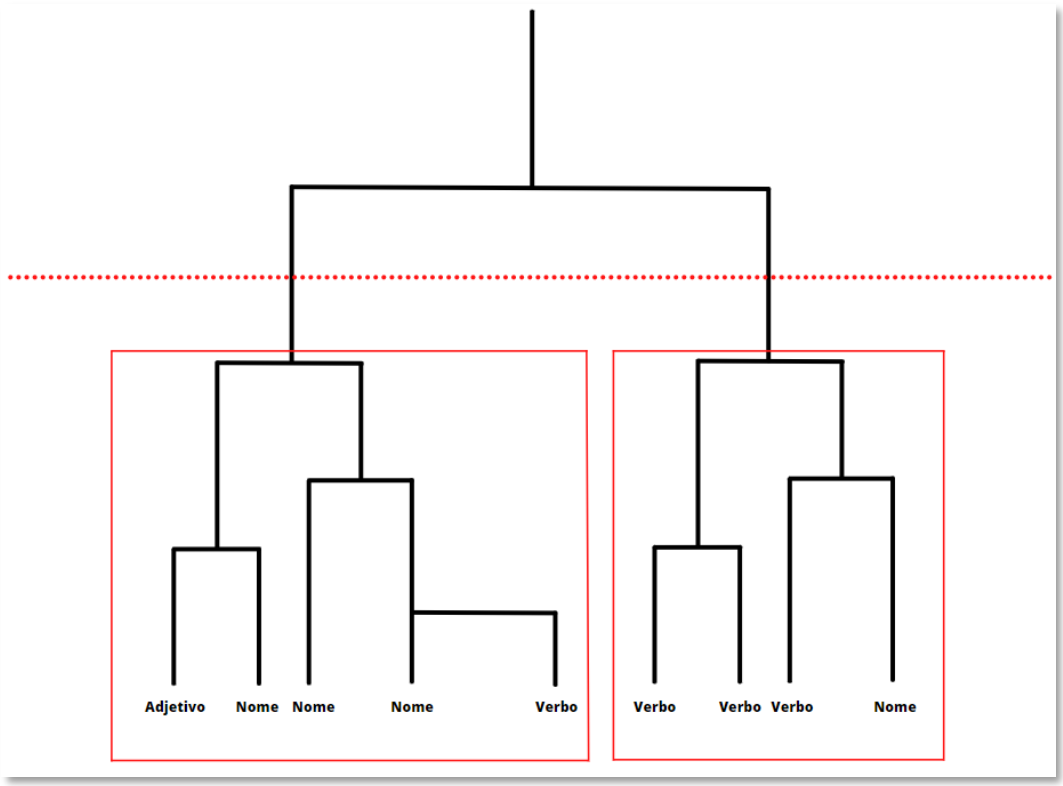
análise de similaridade, foi utilizada a menor distância entre duas linhas da tabela (dois *vetores de contexto*). Essa distância foi calculada a partir do menor ângulo formado entre esses vetores[14]. Assim,

[13] Conforme os autores relatam, não há uma medida única para a similaridade e ela pode ser diferente para cada estudo. Para maiores informações, consultar Redington *et al.* (1998).

[14] Para maiores informações sobre o cálculo utilizado para o ângulo, consultar Mintz *et al.* (2002).

o menor ângulo indica a maior similaridade entre as palavras. Após o cálculo, o resultado foi submetido à *análise hierárquica de cluster* (HCA), que é uma representação hierárquica de melhor ajuste de similaridade de todos os pares de palavras. Para avaliação da performance, alternativamente à medida de *informatividade* utilizada em Redington et al. (1998), Mintz et al. (2002) propõem uma medida denominada de *pureza*, que indica o quão puro um cluster é, ou seja, o quanto um *cluster* agrupa apenas palavras de uma determinada categoria sintática. Essa medida é utilizada para identificar a melhor linha de corte na hierarquia de *clusters*, de modo a obter o agrupamento mais preciso.

Figura 1: Representação de um dendrograma em que o corte é representado pela linha tracejada, gerando dois grandes grupos.



Os experimentos apresentados em Redington *et al.* (1998) demonstram que em todos os cenários a informação distribucional se demonstrou útil e acima da medida base, conseguindo uma precisão média de 72% e completude média de 47%, contra 27% de precisão e 17% de cobertura do piso. Na variação do contexto, foi possível observar que as posições mais próximas às palavras se mostraram mais úteis para a categorização que as mais distantes, mas que a combinação dessas duas posições se mostrou mais efetiva. Na variação de palavras de contexto e palavras-alvo, mesmo conseguindo resultados acima do piso, quando a janela se torna muito grande, o resultado se aproxima do resultado do piso. Analisando separadamente o desempenho nas categorias sintáticas,

os nomes se mostraram mais fáceis de serem categorizados, com precisão de 90% e completude de 53%. Variando o tamanho do *corpus*, o modelo foi se comportando melhor a medida em que a quantidade de palavras foi aumentando, saindo de 60% de precisão e 1% de cobertura, com 500 mil palavras, para 72% de precisão e 42% de cobertura com 1 milhão de palavras. Ao marcar os enunciados explicitamente, houve uma pequena melhora no método, mas não o suficiente para colocar em xeque o desempenho padrão, segundo os autores. Ao verificar se o modelo é sensível a frequência ou apenas a ocorrência das palavras, o resultado padrão, utilizando a frequência e o coeficiente de *Spearman* se mostrou melhor. Quando o modelo é analisado sem as palavras funcionais, os resultados são piorados significativamente, se comparado com a versão base, mostrando que essas palavras são úteis para o método utilizado. Menos do que remover palavras funcionais, mas ao substituir as palavras pela sua determinada categoria faz com que o método tenha um desempenho pior do que o padrão. E, por último, ao utilizar um *corpus* geral e não um *corpus* de fala dirigida à criança, o modelo se mostrou um pouco melhor.

Analisando os experimentos apresentados por Mintz (2002), ao variar a janela de contexto, tanto para os nomes quanto para os verbos, a janela mais informativa foi a com 8 palavras de cada lado da palavra alvo, onde alcançaram 77% de pureza para os nomes e 71% para os verbos. Utilizando a limitação de palavras de classe fechada para o contexto, na categorização dos nomes, a janela mais informativa foi a com 1 palavra ao redor da palavra alvo, com 78% de pureza e, para os verbos, duas janelas obtiveram o mesmo valor médio de 79%, a janela com 2 e 8 palavras. E substituindo as palavras de função pelo símbolo FNCT, na média da pureza, a janela de contexto com melhor resultado foi a com 8 palavras, obtendo uma pureza de 76% para nomes e 83% para verbos. Tais resultados evidenciam a possibilidade de a análise distribucional ser parte importante dos mecanismos da aquisição de linguagem. Os autores ressaltam, ainda, que outras fontes de informação podem contribuir para uma melhor categorização dessas palavras (sintática, fonológica, prosódica ou semântica), ao usar palavras de classes fechadas para limitar o contexto das palavras.

Retomando os conceitos de Harris (1954) apresentados na seção *1.2*, com o modelo apresentado acima, temos que o *Elemento* é o modo como dividimos a entrada para processamento com a ideia de categorização das palavras. Essa entrada pode ser processada como sentenças separadas (cada uma, uma lista ordenada de palavras), respeitando a pontuação ou o texto como uma única sentença, por exemplo. A *Similaridade* é a medida que é utilizada para verificar o quão próximo dois vetores estão um do outro, por exemplo, a medida de correlação de *Spearman*. A *Dependência* pode ser considerada como a própria análise do contexto, porém, sem a necessidade estrita da busca por bigramas, mas sabendo que a coocorrência desses bigramas são úteis para o contexto, por exemplo, a sequência *Artigo* + *Nome*. Não buscamos por isso exclusivamente, mas a

repetição desse padrão nos permite um contexto mais previsível. A *Substitutibilidade* é um dos critérios utilizados pelos modelos que utilizam a análise de contexto para verificar se uma palavra pertence à mesma categoria sintática da outra. Já os *Domínios*, que são apresentados de forma mais robusta em Harris (1954), consideramos, para o modelo aqui proposto, que seriam os contextos em que as palavras ocorrem, desse modo por ser uma simplificação quando comparado com o do autor, a relação estrutural não é representada, mas pode ser um fator que auxilia o modelo. Por fim, os *Dados* no modelo são os *corpora* utilizados para o processamento das informações.

3.2 MODELANDO A MORFOLOGIA

Uma das dificuldades encontradas quando modelamos é representar computacionalmente o fenômeno do mundo real a ser estudado, como apontamos anteriormente. Com a modelagem da morfologia não é diferente pois, como vimos, apesar de haver uma grande quantidade de estudos sobre aprendizagem distribucional, apenas em dois deles (CLARK, 2003; ONNIS e CHRISTIANSEN, 2008), houve inclusão da informação morfológica. No entanto, optamos por nossa própria estratégia de modelagem da morfologia, em face das particularidades do nosso modelo. Para isso, nos baseamos nas propriedades e análises morfológicas encontradas em diversas fontes, tais como Ribeiro (2009), fazendo adaptações que acreditamos serem necessárias. Portanto, nesta seção, explicamos as decisões que foram tomadas e suas justificativas, além das principais dificuldades e de como adaptamos a morfologia para os experimentos. Primeiramente, tomemos como exemplo possíveis análises, apresentadas em Ribeiro (2009)[15]:

(1)

Palavra	**R**	**VT**	**SMT**	**SNP**
Cantarias	cant-	-a-	-ria-	-s

(2)

Palavra	**R**	**VT**	**SMT**	**SNP**
Estudássemos	estud-	-a-	-sse-	-mos

(3)

Palavra	**MD**	**R**	**DNG**	**DNN**
Garotas	-	garot-	-a-	-s

[15] Siglas utilizadas: R- Radical; VT- Vogal temática; SMT- Sufixo modo temporal; SNP- Sufixo número pessoal; DNG- Desinência Nominal de Gênero; DNN- Desinência Nominal de Número; MD- Morfema Derivacional.

(4)

Palavra	**MD**	**R**	**DNG**	**DNN**
Desleal	des-[16]	-leal	-	-

O intuito de trazer tais exemplos é chamar a atenção para as possibilidades de modelagem. Se o objetivo for inserir novas variáveis no modelo, levando em consideração o que foi discutido até o momento, é necessário avaliar quais informações serão inseridas e como serão feitas. No caso das *pistas morfológicas*, isso significa decidir quais aspectos serão inseridos: (i) no caso do PB, utilizaremos sufixos, prefixos ou ambos?; (ii) iremos segmentar exaustivamente as palavras? Essas são apenas algumas das perguntas que podemos levantar acerca do problema da modelagem da morfologia. Vamos analisar o impacto de cada uma dessas perguntas, olhando para o modelo e observando como ele se comportaria. Precisamos lembrar que cada palavra dentro do modelo é um vetor que contém as ocorrências do contexto, então precisamos transformar a informação morfológica em uma representação semelhante.

Iniciando com a pergunta (i) *utilizaremos sufixos, prefixos ou ambos?* Aqui, a decisão tomada influenciará para onde vamos olhar na palavra, se para o começo ou o final. Se decidirmos olhar apenas para o prefixo, só teríamos a marcação em *desleal* (3) do prefixo *des-*, por exemplo, tendo 0 ocorrências nas demais palavras. Caso escolhêssemos apenas o sufixo, não teríamos a marcação em *desleal* (3) e caso optássemos pela marcação de ambos, todas as palavras receberiam a marcação de acordo com sua análise. Em seguida, entraríamos na pergunta (ii) *vamos segmentar exaustivamente?* E aqui já começamos a separar os caminhos para onde queremos que o modelo siga, se comportando de uma forma diferente, dependendo do que será selecionado.

Supondo que optemos por uma análise morfológica exaustiva e que vamos decompor de acordo com o apresentado nos exemplos (1), (2), (3) e (4), teríamos várias colunas, representando cada uma dessas novas variáveis para cada palavra, independente da categoria sintática. As colunas também seriam aplicadas para cada tipo de vogal temática, de sufixo modo temporal e assim para cada uma das novas variáveis, conforme *Tabela 4*. Em contrapartida, caso optemos por separar apenas radicais de informação afixal, teríamos uma coluna para cada aglomerado de afixos encontrado, como apresentado na *Tabela 5*. Aqui, já conseguimos ver o impacto no modelo a depender da escolha que fazemos.

[16] Apesar de não mudar a categoria, o prefixo pode servir como dica para a categorização da palavra.

Tabela 4: Exemplo de um vetor levando em consideração os sufixos e suas funções, apresentados em (1), (2), (3) e (4).

Palavra	VT -a	SMT -rias	SNP -ssemos	DNG -a	DNN -s	MD -des
Cantarias	1	1	0	0	0	0
Estudássemos	1	0	1	0	1	0
Garotas	0	0	0	1	1	0
Desleal	0	0	0	0	0	1

Tabela 5: Exemplo de tabela caso fossem utilizados apenas os sufixos e prefixos concatenados.

Palavra	-rias	-ssemos	-as	-des
Cantarias	1	0	0	0
Estudássemos	0	1	0	0
Garotas	0	0	1	0
Desleal	0	0	0	1

Precisamos sempre ter em mente a *Tabela 4*. Poderíamos olhar apenas para uma dessas informações morfológicas e não modelar as outras ou, ainda, tratar o problema de outra forma, poderíamos fazer essa manipulação até conseguirmos encontrar a forma que acreditamos ser a melhor para o problema que queremos analisar. E esse é um ponto importante: se trata da escolha mais conveniente e viável – e, se possível, a linguisticamente mais interessante – para o problema que queremos modelar.

Dito isso, o que queremos observar aqui é o impacto da morfologia na categorização de palavras utilizando a informação distribucional. Portanto, os aspectos morfológicos modelados foram pensados tendo em vista um aprendiz ingênuo[17], isto é, que não é sensível a todos os aspectos linguísticos em jogo. Para isso, optamos por uma modelagem conservadora e sóbria, que utiliza o que consideramos o mínimo de morfologia como pista para esta etapa.

Para iniciar essa investigação, a decisão que tomamos foi **utilizar apenas informação sufixal como pista morfológica** e sem segmentação exaustiva. Assim, por exemplo, o SMT e SNP são mantidos como um só elemento, assim como DNG e DNN são tidos como um só. Essa junção dos sufixos foi utilizada para todas as categorias sintáticas. Podemos observar como fica a decomposição da palavra *estavas* em (5), ao invés de uma possível representação como em (6):

(5)

Palavra	**Sufixo**
estavas	-vas

(6)

Palavra	**SMT**	**SNP**
estavas	-va	-s

[17] A referência a ingenuidade foi feita pensando em um aprendiz, que tem todo aparato disponível para as análises e tomadas de decisões.

Também optamos por fazer simulações sem e com a informação sobre a função do morfema. Uma vez decidido isso, o próximo passo foi determinar como inserir tal informação no modelo. Primeiramente, cabe explicar como a extração/seleção do sufixo foi feita. A partir da lista das mil palavras alvo mais frequentes, nossa decisão foi agrupar palavras com base na semântica do radical, por exemplo, em (7):

(7) Menino/menina/meninos/meninas/menininho

Para cada agrupamento deste tipo foi criado um identificador para permitir associar sufixos às palavras correspondentes, conforme apresentado na *Tabela 6*. O identificador é uma representação arbitrária utilizada neste estudo. Vale ressaltar, que a forma do identificador não importa, pois as formas dos radicais por si não entram na análise que o modelo faz, mas apenas os sufixos.

Tabela 6: Exemplo de agrupamento feito no modelo.

Palavra	Identificador	Sufixo
Menino	Men	-o
Menina	Men	-a
Meninos	Men	-os
Meninas	Men	-as

Como é possível observar, o identificador do grupo não coincide necessariamente com a forma do radical; ele significa apenas que essas palavras são candidatas a estarem relacionadas entre si do ponto de vista morfológico. Em outras palavras, assumimos que o aprendiz é capaz de fazer tal hipótese, a partir da observação de regularidades nas palavras mais frequentes que encontra no *input*. Voltaremos a este ponto mais adiante. A partir de agrupamentos como o exemplificado acima, as palavras do grupo foram caracterizadas morfologicamente pela ocorrência dos sufixos encontrados no *corpus*, como ilustrado na *Tabela 7*:

Tabela 7: Exemplo de preenchimento dos vetores morfológicos.

Palavra	Sufixos			
	-o	-a	-os	-as
Menino	617	0	0	0
Menina	0	635	0	0
Meninos	0	0	190	0
Meninas	0	0	0	116

Com isso, formamos um *vetor morfológico* do grupo, como disposto abaixo:

Tabela 8: Representação do vetor morfológico por grupo.

Grupo	Sufixos			
	-o	-a	-os	-as
Men	617	635	190	116

Assim, a integração da informação morfológica ao modelo anterior se dá pela concatenação do vetor morfológico de cada palavra com o vetor de contexto apresentado anteriormente na ***Erro! Fonte de referência não encontrada.***. A representação final do vetor é apresentada na *Tabela 9*.

Tabela 9: Representação final do vetor de contexto.

Vetor de contexto					
PA	Vetor Alvo -2	Vetor Alvo -1	Vetor Alvo +1	Vetor Alvo +2	Vetor Morf.

O objetivo do estudo é investigar se esse *vetor morfológico* permitirá um melhor agrupamento de palavras, verificando o quanto a morfologia inserida no modelo impacta na categorização. Retomando agora a assunção de que o aprendiz é capaz de fazer hipóteses morfológicas, para que uma palavra tenha seu sufixo contabilizado e entre para essa lista no modelo, ela precisa atender a um de dois critérios iniciais:

1. Formar um par mínimo com outra palavra da lista de mil palavras mais frequentes
    a. Menino / Menina
    b. Canto / Cantei
    c. Do / Da

2. Ou, se não houver pares mínimos, o possível sufixo precisa estar entre os sufixos mais produtivos, o que no modelo significa ter uma frequência de mais de 10%:
    a. Revista / -a

Dessa forma, diante dos critérios apresentados, algumas decisões foram tomadas e elencadas a seguir:

1. Só foram analisadas para a morfologia as 1000 palavras-alvo mais frequentes;
2. Para a decomposição morfológica, adotamos a proposta apresentada em Ribeiro (2009);
3. Para os casos em que a palavra tenha mais de um sufixo, como em *bonit-inh-a*, optamos pelo sufixo (não analisado, no caso, *-inha* e *-a*) que determina o maior grupo, ou seja, o

radical que tem mais ocorrências no *corpus*, como mostra a Tabela 10. Essa opção reflete nossa assunção de que o aprendiz é sensível aos radicais e capaz de notar aqueles que cobrem um maior número de palavras. A consequência disso é a de que a análise inicial dos sufixos seria não exaustiva, como discutido anteriormente.

Tabela 10: Exemplo de palavra com dois sufixos possíveis.

	Palavra: bonitinha	
	Sufixo 1	**Sufixo 2**
	-a	-inha
raiz	bonitinh	bonit
total	116	617
selecionado	**-inha**	

4. Para palavras que podem ter dois sufixos, em casos de homofonia, os dois são contabilizados na tabela de morfologia, como forma de capturar eventuais hipóteses incorretas feitas pelo aprendiz a partir dos dados (ruído):

Tabela 11: Exemplo de representação de palavras homônimas, conforme análise de Ribeiro (2009) em que o sufixo de terceira pessoa do singular no presente do indicativo é analisado como sendo vazio.

Palavra	Sufixo 1 (nome)	Sufixo 2 (verbo)
conta	-a	-Ø
contas	-as	-s

5. Para palavras cuja raiz aparece apenas uma vez dentre as mil mais frequentes e que apresentam mais de um sufixo possível, optamos pelo sufixo com maior ocorrência dentre as mil palavras analisadas. Por exemplo, em *coz-inh-a*, onde será selecionado o sufixo *-a*;
6. Nos casos em que uma palavra ocorreu sozinha no *corpus*, não sendo possível contrastar com outra, não marcamos o sufixo *-Ø,* como em *rio*;
7. Quando um substantivo contrasta apenas no plural e não no gênero, acrescentamos o morfema vazio para representar o singular e o morfema flexional *s* como sufixo de plural. Por exemplo, em *hora* e *horas*, marcamos o *singular vazio* em *hora* e apenas *s* em *horas*;
8. Nas situações em que uma palavra aparece com contraste de gênero e plural, como em *outro, outra e outros*, acrescentamos os morfemas de gênero e de número, mas não o vazio[18] para representar singular, pois não decompomos os sufixos:

18 Segundo o princípio 4 de Aronoff (2011, p. 17), um morfema pode ser marcado como – Ø, desde que contraste com outro morfema que não seja – Ø.

Tabela 12: Exemplo de uma representação em que uma palavra ocorre sem singular vazio.

Palavra	Sufixo
outro	-o
outra	-a
outros	-os

9. Existem casos limítrofes, nos quais consideramos que o aprendiz suporia uma relação semântica possível em função da similaridade das formas e que, por isso, colocamos no mesmo grupo, como é o caso de *bolo*/*bola* (refletindo, assim também, situações de ruído):

Tabela 13: Exemplo de um caso limítrofe

Palavra	Grupo	Sufixo
bola	bol	-o
bolinha	bol	-a
bolo	bol	-os

10. Verbos irregulares cujas formas flexionadas variam ao ponto de não ter um radical em comum ou que não tenham um radical dentro de um grupo já montado, como os verbos *temos* e *tinha*, participam de grupos diferentes;
11. Outros verbos irregulares ficaram de fora da análise morfológica por não haver outras palavras para contrastar dentre as mil mais frequentes; e
12. Em palavras como *jogo*, em que a forma do sufixo seria a mesma tanto para o substantivo como para o verbo correspondente, mantivemos apenas o sufixo -o.

Em algumas situações, podemos ter radicais ou sufixos que não estão previstos em Ribeiro (2009), ou que não seriam os primeiros a serem escolhidos por uma análise feita por um linguista. Além disso, estamos simplificando para formas ortográficas e não representações fonéticas. Porém, temos noção desses aspectos e, como dito no começo desta seção, a intenção é iniciar com uma modelagem mais ingênua, que nos permita sondar o terreno e aos poucos avançarmos na modelagem da morfologia na direção de algo linguisticamente mais enriquecido. O conjunto final das categorias sintáticas pode ser visto no *Gráfico 1*, em que são apresentadas a quantidade e porcentagem de cada uma dessas categorias no conjunto das 650 palavras. E, no *Gráfico 2*, apresentamos uma comparação com as categorias encontradas nas 1000 palavras mais frequentes.

Gráfico 1: Distribuição da quantidade das categorias sintáticas após a inclusão da informação morfológica, resultando no total de 650 palavras. Ao lado esquerdo do ',' é apresentado o valor total de palavras daquela categoria e a direita a porcentagem que ela representa no *corpus*.

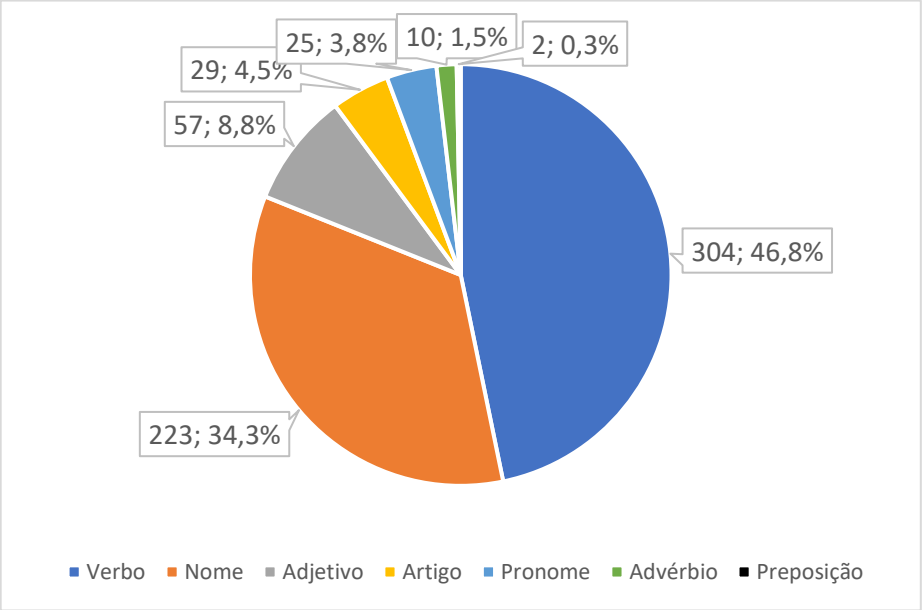
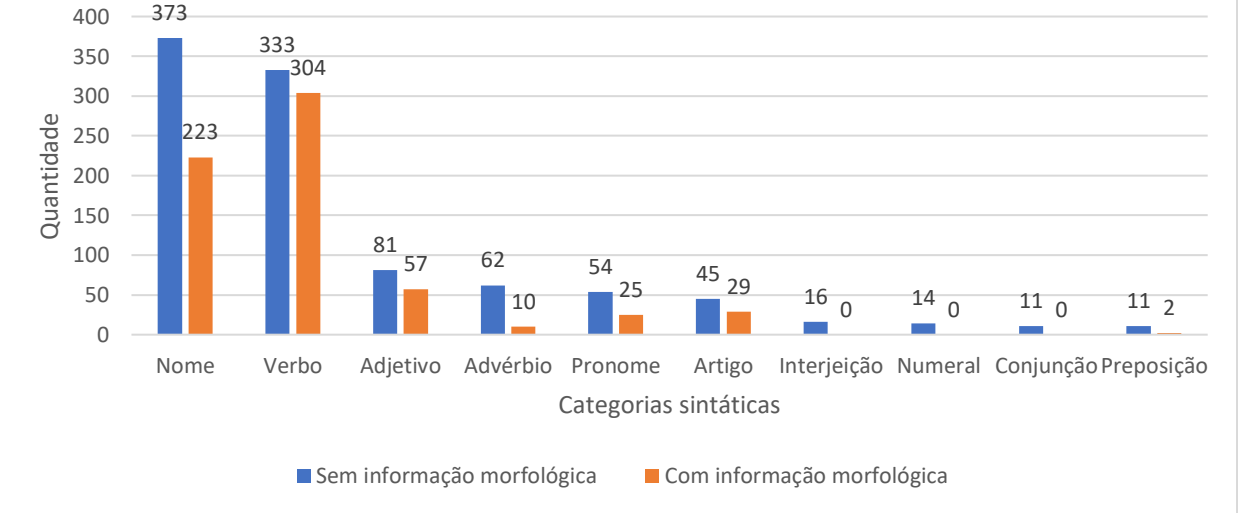


Gráfico 2: Comparação entre a quantidade palavras presentes nas categorias sintáticas com informação morfológica (650 palavras) e sem informação morfológica (1000 palavras).



3.3 OS CORPORA

Os *corpora* consistem em dois conjuntos de dados: de fala dirigida à criança e de diálogos entre adultos. Os primeiros foram compilados, inicialmente, a partir da *Coleção Projeto de Aquisição da Linguagem Oral (IEL/Unicamp)* e dos dados do PB disponíveis na base *CHILDES*. Já o segundo conjunto, foi obtido da plataforma *Projeto Norma Linguística Urbana Culta – RJ (NURC)*, porém, só são utilizados para o conjunto de experimentos base e não para os experimentos que utilizam a

morfologia, esses só utilizaram o *corpus* de fala dirigida à criança. Idiossincrasias, por exemplo, *popó*, não foram classificadas; e somente as pontuações finais foram mantidas e substituídas por ponto final, totalizando mais de 1,43 milhão de palavras. Porém, alguns experimentos precisam de uma abordagem diferente, então adiante, na seção *3.5* foi exibida a configuração de cada experimento feito, inclusive como os *corpora* foram tratados para cada um deles.

3.4 PERFORMANCE

Para avaliar o desempenho do modelo, utilizamos a mesma medida apresentada em Faria e Ohashi (2018), a medida *F*, que é a média harmônica entre a *precisão* e *cobertura*. Para que essas medidas sejam aplicadas, tanto os grupos das palavras de referência, quanto os grupos dos resultados, são separados em pares de palavras. Assim, para a *precisão*, o cálculo leva em consideração a quantidade de pares palavras corretas que estão dentro dos grupos criados, segundo a classificação de referência. Por exemplo, se 50 pares de palavras foram categorizados e apenas 10 pares estiverem corretos, então, a *precisão* para esse exemplo é de 20%. Na *cobertura*, nós temos a informação de quantos pares de palavras deveram estar no mesmo grupo. Ou seja, imaginando que o método deveria supor a existência de 40 pares de palavras corretos, mas só 10 foram criados, temos que a completude foi de 25%. Como temos uma relação de dependência entre as duas medidas, precisamos configurá-las para que o retorno seja de acordo com a nossa necessidade: quanto maior precisão, menor tende a ser a cobertura; e quanto maior a cobertura, menor tende a ser a precisão. Por isso, é essencial saber o que é mais importante para o seu método, se é a *precisão* ou a *cobertura*. A fórmula que rege a medida *F* pode ser vista em (8):

(8) $$F_\beta = (1 + \beta^2) * \frac{precisão * cobertura}{(\beta^2 * precisão) + cobertura}$$

Com o intuito de favorecer a *precisão* e manter as mesmas configurações utilizadas por Faria e Ohashi (2018), o coeficiente *β* foi definido em 0,3. A intenção de priorizar a *precisão* decorre da observação de que a criança é precisa nas suas categorizações. Além disso, essa configuração também é interessante para o equilíbrio do método, uma vez que a cobertura acaba sendo favorecida pela existência de grandes categorias sintáticas como nomes e verbos, ou seja, se o modelo acertar apenas estas, obtemos uma boa cobertura, mas a precisão fica muito comprometida quando pensamos nas categorias funcionais.

3.4.1 Classificação de referência e piso classificatório

Para avaliarmos a *performance* do método, precisamos de uma classificação de referência e de um *piso* classificatório. A classificação de referência é para que possamos verificar se o modelo está acertando no processo de categorização. Já o *piso* classificatório é necessário para que consigamos avaliar o quão melhor – mais informativo – o modelo desempenha, quando comparado a um desempenho aleatório em que as palavras são agrupadas às cegas.

Para a classificação de referência, foi utilizada a mesma do estudo de Faria e Ohashi (2018). O *corpus* para extração das categorias de referência utilizado foi o *Corpus Histórico do Português Tycho Brahe* (CTB), composto por textos do português europeu e brasileiro, editados eletronicamente e etiquetados morfossintaticamente. Por ser um *corpus* histórico, algumas palavras contemporâneas, como *computador* e *televisão*, além de nomes próprios, formas diminutivas, entre outras, tiveram que ser incluídas manualmente. Algumas palavras idiossincráticas, tais como *popó* não foram classificadas, ficando excluídas do processamento.

O método opera com a assunção simplificadora de que a cada palavra corresponde apenas uma categoria, no caso, a categoria mais frequente no *corpus* para cada uma. Portanto, se a palavra *poder* aparecer como verbo 80% das vezes no *corpus*, essa será a categoria assumida pelo modelo. Ademais, para que fosse possível utilizar as mesmas etiquetas adotadas por Redington *et al.* (1998), o padrão de etiquetas do CTB foi convertido para uma versão reduzida. Para etiquetas que envolvem combinações de duas ou mais, como P+D (pronome + determinante), por exemplo, apenas a primeira foi considerada. Um resumo dessa conversão é apresentado na *Tabela 14*.

Tabela 14: Exemplo de conversão utilizada do padrão CTB para a simplificação utilizada em Redington (1998).

Categoria	Etiquetas do CTB	Exemplo	Frequência
Substantivo	N, NPR	ademir, adriana, ajuda	375
Adjetivo	ADJ, OUTRO	alto, amarelo, baixo	82
Numeral	NUM	cinco, dez, duas	14
Verbo	VB, HV, ET, TR, SR	abre, abrir, abriu	331
Artigo	D	a, aquele, os	45
Pronome	CL, SE, DEM, PRO, PRO$, SENAO, QUE, WADV, WPRO, WD,WPRO$, WQ	aonde, aquilo, cadê	53
Advérbio	ADV, Q, NEG, FP	agora, ainda, algum	62
Preposição	P	até, com	11
Conjunção	CONJ, CONJS, C	como, e, enquanto	11
Interjeição	INTJ	ah, ahn, ai	16

Fonte: Adaptada de Faria e Ohashi (2018).

O *piso* classificatório é gerado da mesma forma que em Redington *et al.* (1998): para cada nível de similaridade, os grupos são mantidos e palavras são redistribuídas de forma aleatória entre os grupos e a *performance* é calculada. Esse procedimento é repetido por dez vezes e então calcula-

se a média da *performance*. A partir dessas duas classificações, a de referência e a aleatória, conseguimos avaliar o resultado obtido pelo modelo em cada experimento.

3.5 OS EXPERIMENTOS

Os experimentos apresentados no capítulo *4 Resultados e discussão* são aqueles conduzidos em Redington *et al.* (1998) e Faria (2019). Ao todo, são 9 experimentos, em que cada um visa a investigação de um determinado aspecto. Em cada experimento, são executadas uma ou mais simulações com diferentes parâmetros, totalizando 49 simulações. Os 9 experimentos são: (1) variação da janela de contexto; (2) variação da quantidade de palavras alvo e de contexto; (3) avaliação do método para cada classe de palavras; (4) variação do tamanho do *corpus*; (5) mudança na fronteira do enunciado; (6) variação na medida de distância; (7) remoção das palavras funcionais; (8) substituição da palavra por sua classe; e (9) utilização do *corpus* de fala entre adultos.

Dentre essas várias simulações, o experimento *standard* consiste em analisar as 1000 palavras mais frequentes como palavras-alvo utilizando o *corpus* de fala dirigida à criança. Como palavras de contexto válidas, são utilizadas as 150 mais frequentes, com uma janela que cobre as duas palavras anteriores ao da palavra-alvo e duas palavras seguintes, representada como [-2, -1, 1, 2]. Como são 4 posições a serem consideradas e temos 150 palavras de contexto possíveis, um vetor de contexto de 600 posições é formado para cada palavra-alvo, visto que a estatística de cada palavra de contexto precisa ser contabilizada para cada uma das posições de contexto possíveis. Para esse experimento, a pontuação é retirada e o *corpus* é concatenado, formando uma única sentença, somando um total de 1,15 milhão de palavras. Como todas as outras configurações são variações do experimento *padrão*, caso algum parâmetro não seja mencionado, ele foi mantido como no *padrão*. A seguir, apresentamos como cada experimento foi configurado seguindo o que foi apresentado em Redington *et al.* (1998):

1. Experimento 1 – Diferentes contextos: para a categorização de palavras utilizando a informação distribucional assumimos que as crianças são sensíveis ao contexto, assim, precisamos saber quais janelas de contexto são mais informativas no processo de categorização e se essas janelas são possíveis de serem detectadas facilmente pela criança. Então, esse experimento tem o objetivo de verificar qual é a janela de contexto que mais auxilia a categorização das palavras. Para isso, a janela de contexto é variada, verificando algumas condições, verificando

contextos individuais a esquerda ou direita das palavras-alvo e verificando os dois lados simultaneamente.

2. Experimento 2 – Variando o número de palavras alvo e de contexto: Qual a quantidade de palavras-alvo e de contexto devemos usar para que o método utilizado seja eficaz e é realista para uma criança? Para ajudar a entender essa pergunta, variamos a quantidade dessas palavras mantendo as outras configurações sem variação.
3. Experimento 3 – Para quais categorias o método funciona melhor: a fim de verificar se a informação distribucional de contexto auxilia as categorias de forma diferente, fazemos uma análise de desempenho por categoria sintática, na qual a configuração do experimento padrão é utilizada, medindo o grau de informatividade para cada uma.
4. Experimento 4 – Tamanho do *corpus*: a quantidade de palavras em que a criança é exposta já pode ser utilizada para justificar o tamanho do *corpus* utilizado, porém, para verificar o comportamento do método em casos com uma quantidade menor de palavra, fazemos testes em que a quantidade total é diminuída.
5. Experimento 5 – Fronteira das sentenças: na análise base, todas as palavras são concatenadas e o *corpus* vira uma longa sentença, isso pode tanto melhorar quanto piorar o método. Com o intuito de verificar o impacto das marcações das sentenças e aproximar mais da realidade da criança, fizemos duas análises, uma em que a análise acontece apenas na sentença e uma segunda, onde a marcação da fronteira é marcada de forma explícita. Assim, conseguimos avaliar o impacto das fronteiras do enunciado.
6. Experimento 6 – Frequência *vs.* Ocorrência: o experimento base assume que as crianças são sensíveis a frequência de palavras, o que é plausível. Porém, é interessante analisar o que acontece quando o método substitui a análise de frequência por ocorrências, sendo assim, fazemos 3 experimentos que analisam essas situações.[19]
7. Experimento 7 – Removendo palavras funcionais: analisa o fato das crianças prestarem mais a atenção nas palavras de conteúdo do que nas palavras funcionais, assim avalia o impacto das palavras funcionais no método. Para isso todas as

[19] Nesse experimento é testado um método de distância diferente, o city-block em que a distância entre dois pontos é a soma das diferenças absolutas de suas coordenadas.

palavras funcionais são removidas do *corpus*, possibilitando verificar como se o aprendiz não tivesse esse *input*.

8. Experimento 8 – Verificar se as informações sobre uma categoria ajudam na aquisição de outras: as crianças, provavelmente, utilizam outras pistas no processo de categorização. Então, é interessante perguntar até que ponto outras dicas ajudam nesse processo. Aqui, verificamos se o conhecimento sobre uma determinada categoria impulsiona a categorização de outras categorias, substituindo as palavras por suas respectivas categorias em cada um dos experimentos, por exemplo, todos os nomes sendo substituídos por NOUN, todos os verbos por VERBS.
9. Experimento 9 – Verificar se o aprendizado com fala dirigida à criança é mais fácil: o objetivo aqui é verificar se a linguagem dirigida à criança contribui mais para a aquisição, por ser diferente em termos de vocabulário e complexidade sintática da fala que não é dirigida a criança. Então, observamos o impacto do *corpus* de fala entre adultos no processo de aquisição, substituindo o *corpus* de fala dirigida à criança, pelo *corpus* do NURC.

4 RESULTADOS E DISCUSSÃO

Alguns dos resultados obtidos aqui (cf. item 4.1), aqueles dos estudos sem morfologia, já foram discutidos em Faria e Ohashi (2018), Faria (2019a) e Faria (2019b). É importante observar que algumas mudanças no código ou na base de dados, mesmo que pequenas, podem gerar diferenças nos resultados. Portanto, os resultados das várias condições experimentais na condição sem morfologia relatados adiante, não necessariamente reproduzem exatamente os citados em artigos anteriores. Em geral, porém, estas diferenças são pouco significativas e não mudam o quadro geral. A versão final do código-fonte do presente modelo será disponibilizada em um repositório público, no futuro.

Os experimentos apresentados aqui são derivados do que irei chamar de *simulação base*, que seguem as configurações propostas em Redington *et al.* (1998). O experimento 9 consiste na fala entre adultos, então não entram nas simulações que utilizam a morfologia, uma vez que elas não utilizam esse *corpus*. Sendo assim, as simulações são apresentadas seguindo as seguintes divisões:

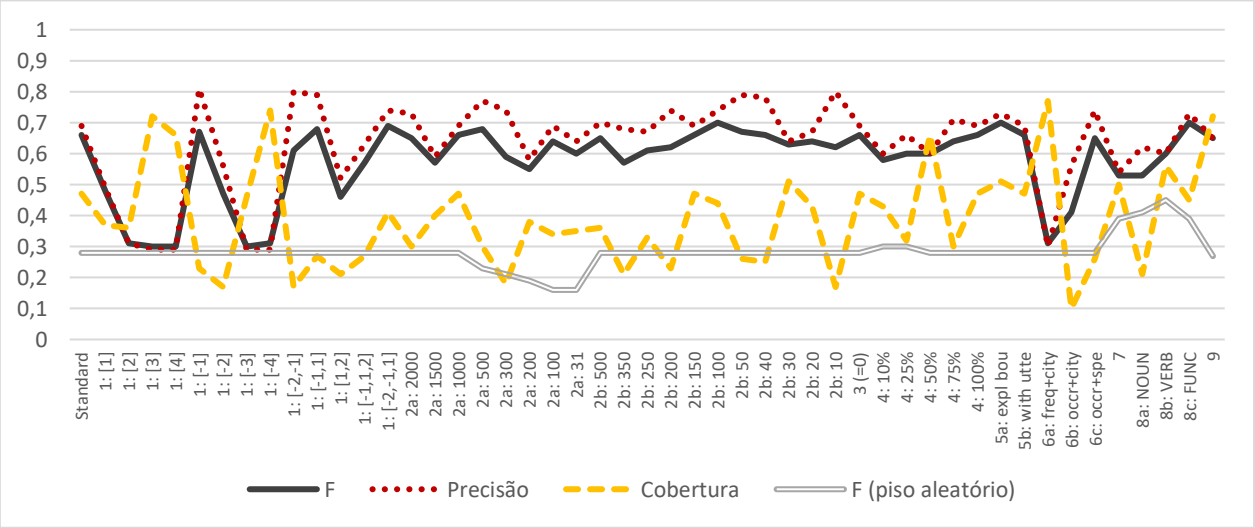
- Simulação base: as nove condições experimentais implementadas na versão do modelo *sem* morfologia;
- Simulação 1: simulação base acrescida das informações morfológicas;
- Simulação 2: simulação base acrescida das informações morfológicas e das funções morfológicas;
- Simulação 3A: simulação base restrita às 650 palavras selecionadas para receber as informações morfológicas, mas sem o uso dessas informações;
- Simulação 3B: simulação 1 restrita às 650 palavras e com as informações morfológicas;
- Simulação 3C: simulação 2 restrita às 650 palavras com a informação morfológica e com as suas respectivas funções morfológicas;
- Simulação 4: restrita às 650 palavras e usando apenas as informações da morfologia, sem utilizar a informação distribucional do contexto.

Inicialmente, apresentamos uma análise quantitativa dos dados para todas as simulações e finalizamos com uma análise qualitativa da condição *standard* da *simulação 3B*.

4.1 SIMULAÇÃO BASE

Os resultados apresentados nesta seção serviram como base de comparação para as demais simulações do modelo. No *Gráfico 3*, podemos observar o comportamento do modelo para todos as condições e suas variações, totalizando 49 condições experimentais. No gráfico, é possível ver a precisão, a cobertura, o melhor *F* e o melhor *F* do piso aleatório, para todos as condições, ficando evidente, exceto por cinco condições experimentais, que o aprendiz sempre obtém um valor de *F* significativamente maior do que o *piso aleatório*. As condições com piores desempenhos foram aquelas em que a janela de contexto considerou apenas uma palavra nas posições [2], [3], [4], [-3] e [-4], e na condição em que utilizamos a métrica de similaridade *city-block (condição "6a: freq+city")*.

Gráfico 3: Resultado geral da *simulação base*, com o melhor resultado encontrado para cada condição experimental, com sua precisão e cobertura.



No *Gráfico 3* observamos os melhores resultados das condições experimentais, descritas na seção *3.5*. A variação do contexto é representada pelo número que está entre os colchetes, por exemplo, a configuração *1:[1]* indica que é o experimento 1 e que a janela que está sendo analisada é a janela imediatamente seguinte e verificamos posição à esquerda e à direita simultaneamente, sendo representada por *1:[-1,1]*, por exemplo. O experimento 2 é divido em variações de palavras-alvo e variações de palavras de contexto, sendo que o experimento *2a:2000* indica que é variação de palavras alvo (a) e que está utilizando 2000 palavras. Já o *2b:500*, indica que é o experimento variando o contexto (b) e que está utilizando 500 palavras. No experimento 4 temos a representação da parcela utilizada do *corpus* pela porcentagem, onde o experimento *4:10%* indica que estamos utilizando 10% do *corpus* total. O experimento 5 se divide em duas configurações, o *5a: expl bou*,

sentenças separadas e em que a fronteira das sentenças recebe as marcações de forma explícita; *5b: with utte* em que a análise é restrita à sentença, mas sem fronteiras explícitas. As configurações do experimento 6 consistem em 3 condiçõe:s, a primeira, representada por *6a: freq.+city* é o experimento onde é utilizada a frequência das palavras com a métrica *cityblock*; o *6b: occr+city*, que também utiliza a *cityblock*, mas agora apenas com a ocorrência das palavras; e, por último, uma configuração com a ocorrência das palavras e com *spearman*, *6c: occr+spe*. Já para a o experimento 8, a categoria sintática é explicitada logo após o número do experimento, *8a: NOUN*, que é o experimento que faz as substituições nos nomes, *8b: VERB*, substitui o verbo e *8c: FUNC* substitui as palavras funcionais.

Os melhores resultados encontrados nessa simulação podem ser vistos na *Tabela 15*. As discussões sobre esses experimentos para a simulação base já foram feitas em Faria e Ohashi (2018), Faria (2019a) e Faria (2019b).

Tabela 15: Melhores resultados da *simulação base*.

Experimento	Melhor F	Precisão	Cobertura
Standard	0,66	0,69	0,47
1:[-2,-1,1]	0,69	0,74	0,41
2b: 100	0,7	0,74	0,44
3:noun	0,72	0,72	0,67
4: 100%	0,66	0,69	0,47
5a: expl. bou	0,7	0,73	0,51
6c: occr + spe	0,65	0,74	0,26
7	0,53	0,54	0,5
8c: FUNC	0,7	0,73	0,45
9	0,65	0,65	0,72

4.2 SIMULAÇÕES COM MORFOLOGIA

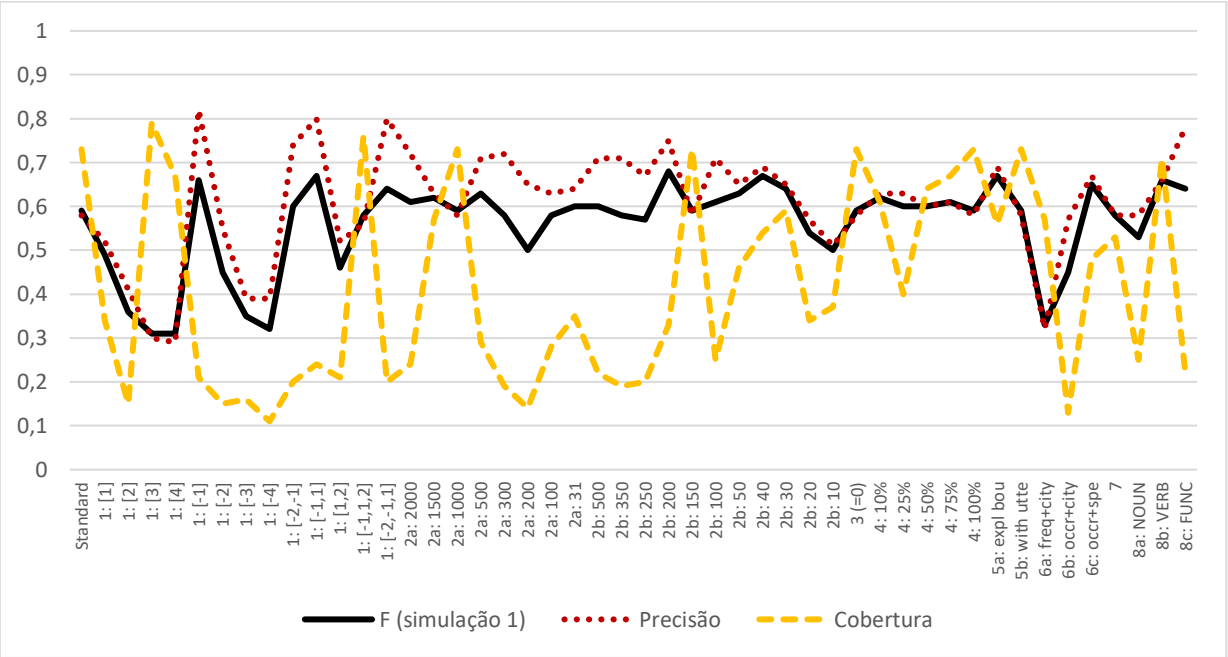
As condições experimentais apresentadas nesta seção são as mesmas feitas na *simulação base*, mas com a informação morfológica inserida nas 650 palavras que foram selecionadas segundo nossos critérios, as demais permanecendo sem análise (o que produz, na prática, um vetor morfológico zerado para elas.

4.2.1 Simulação 1: utilizando a informação morfológica

No *Gráfico 4*, observamos os resultados da medida *F*, além da *cobertura* e *precisão* para todas as condições dos experimentais. Assim como na simulação base, os piores resultados se mantiveram nas mesmas condições, onde analisamos uma palavra como janela de contexto nas

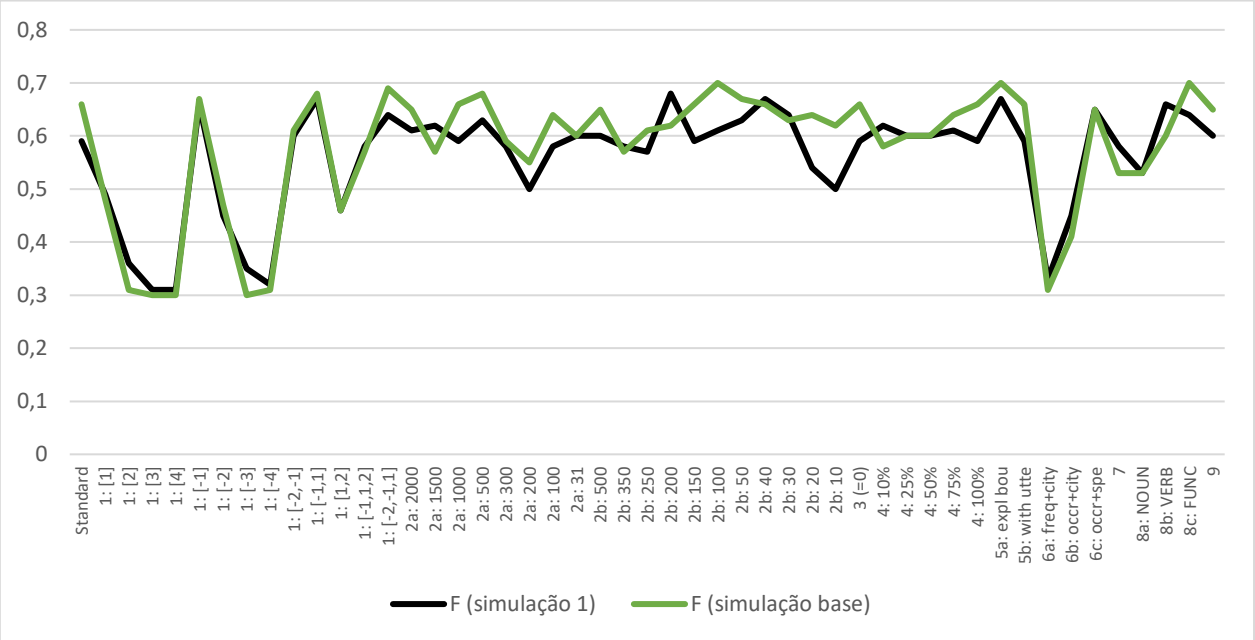
posições [2], [3], [4], [-3] e [-4], e na condição em que utilizamos a métrica de similaridade *cityblock (condição "6a: freq+city").*

Gráfico 4: Resultados com a morfologia para todos as condições experimentais da *simulação 1*.



No *Gráfico 5* apresentamos um contraste entre a *simulação base* e a *simulação 1* para todas as condições experimentais. Comparando diretamente com o que foi obtido na Simulação Base, temos uma piora no modelo em várias condições, inclusive na *standard*. Temos uma hipótese sobre isso. Vale lembrar que o vetor morfológico é concatenado ao vetor distribucional. Ocorre que, para palavras que não receberam análise morfológica segundo nossos critérios, seus vetores morfológicos contêm apenas zeros. Como o método considera um vetor zerado como informação, isso cria um ruído no modelo, pois na prática é como se ele considerasse todas estas palavras como sendo morfologicamente iguais. Uma observação importante deve ser feita quanto às condições experimentais que variam a quantidade de palavras-alvo. Como as 650 palavras que receberam análise morfológica estão distribuídas dentre as 1000 mais frequentes, as condições que variam entre 31 e 500 palavras-alvo não necessariamente incluem apenas palavras morfologicamente analisadas, pois esta seleção se dá apenas pela frequência, sem levar em consideração haver ou não análise morfológica. Os melhores resultados para a *simulação 1* podem ser vistos na *Tabela 16*.

Gráfico 5: Resultados comparando o F da simulação 1 e da simulação base.



É possível identificar algumas diferenças se comparados com a Simulação base. Por exemplo, para a condição variando a janela de contexto, o melhor resultado ocorre usando a janela [-1,1] com um *F* de 0,67, já o base alcança um *F* de 0,68 para a mesma janela. Sendo que o melhor *F* alcançado é utilizando a janela [-2,-1,1], com um *F* de 0,69, demostrando uma piora no desempenho. Na condição variando a quantidade das palavras de contexto, obtivemos um resultado melhor com 200 palavras, ao invés de 100, como na *simulação base*. Quanto ao tamanho do *corpus*, o melhor resultado foi obtido com 10% do *corpus*, enquanto no base o melhor foi com 100%. E, na condição 8, ao substituir as palavras de função por VERB, ao invés de FUNC, temos o melhor resultado. Essas diferenças podem ser explicadas pelo que foi dito antes, a falta de informação em algumas palavras faz com que o método se confunda em algumas situações, mas melhore em outras em que mais palavras etiquetadas sejam utilizadas.

Tabela 16: Melhores resultados da *simulação 1*.

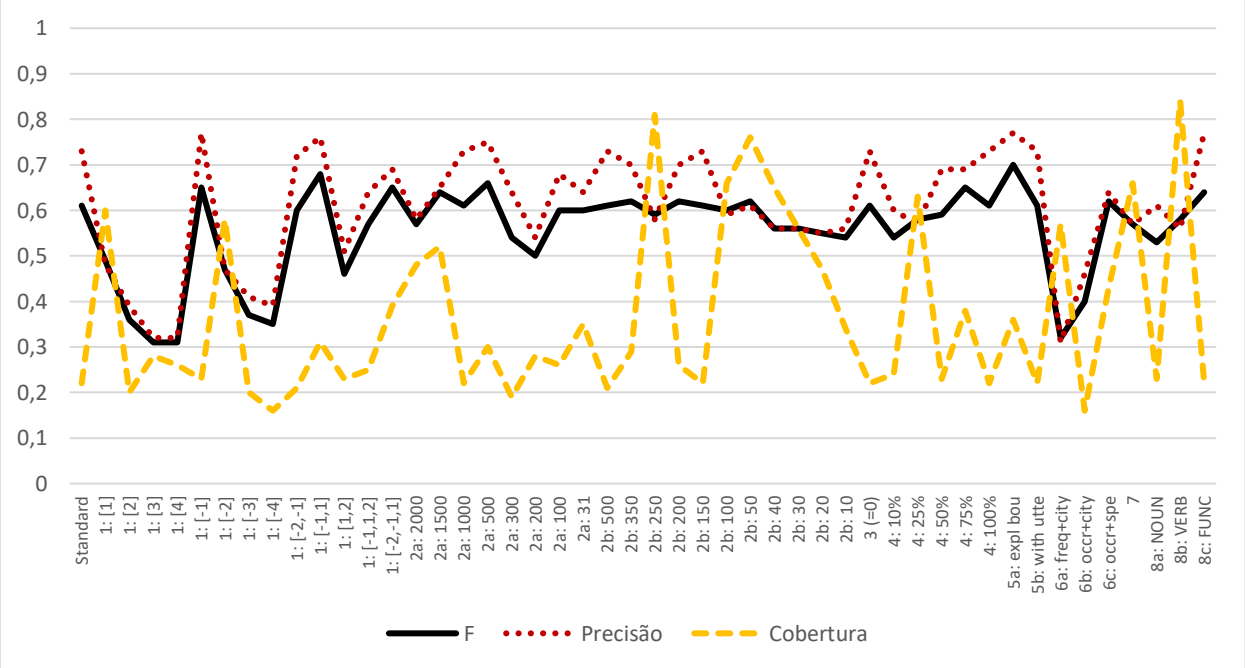
Experimento	Melhor F	Precisão	Cobertura
Standard	0,59	0,58	0,73
1: [-1,1]	0,67	0,8	0,24
2b: 200	0,68	0,75	0,33
3:noun	0,72	0,71	0,73
4: 10%	0,62	0,63	0,61
5a: expl bou	0,67	0,69	0,56
6c: occr+spe	0,65	0,67	0,48
7	0,58	0,58	0,53
8b: VERB	0,66	0,66	0,71

Em um primeiro momento, podemos nos surpreender com os resultados encontrados nessa simulação. No entanto, em função dos vetores morfológicos zerados para palavras não analisadas, é possível que não tenhamos conseguido mensurar, de fato, o impacto da morfologia no aprendizado. Isso porque, no cômputo da performance, todas as palavras alvo, com e sem morfologia, são consideradas. Se os vetores zerados funcionarem como ruído, como hipotetizamos, isso vai prejudicar a performance geral. Outra possibilidade, não exclusiva, é a de que a informação morfológica e a distribucional possam estar conflitando. Para investigar tudo isso, resolvemos criar outras simulações utilizando apenas as palavras-alvo com análise morfológica. Os resultados são apresentados na seção *4.3*, em que conseguimos avaliar o impacto da inserção de informação morfológica. Antes disso, porém, apresentamos a simulação em que utilizamos também a função morfológica como informação que distingue entre formas sufixais.

4.2.2 Simulação 2: utilizando a informação morfológica e a função

As mesmas análises que fizemos na *simulação 1*, fizemos aqui. As piores condições experimentais são as mesmas já relatadas, em que é analisada uma palavra como janela de contexto nas posições [2], [3], [4], [-3] e [-4] e na condição em que utilizamos a frequência do contexto com o cálculo de distância *cityblock (condição "6a: freq+city")*. Uma visão geral da simulação pode ser vista no *Gráfico 6*.

Gráfico 6: Resultados com a morfologia e função para todos as condições experimentais da *simulação 2*.



Comparando os melhores *F* obtidos nas simulações, podemos ver no *Gráfico 7* que o modelo com forma e função se comportou pior que o modelo base e o modelo utilizando apenas a forma. Como antes, terminamos com a conclusão contraintuitiva de que a informação morfológica não agrega informação e até atrapalha a aprendizagem distribucional. Como estamos inserindo a função dos sufixos também, isso faz com que o tamanho do vetor morfológico aumente, porque se antes o modelo tinha apenas o *-a* como possível sufixo, agora ele tem as variações, onde *-a* pode ser um sufixo de gênero para os nomes ou um sufixo de flexão para os verbos (considerando a morfologia simplificada assumida na modelagem). Desse modo, o "ruído" das 350 palavras sem análise pode estar causando ainda mais dificuldades para o aprendiz.

Os melhores resultados das simulações podem ser vistos na *Tabela 17*, onde comparamos os resultados encontrados em todas as simulações: *simulação base*, *simulação 1*, *simulação 2*. Podemos ver que os melhores resultados foram alcançados pela *simulação base*. De todo modo, apesar da informação morfológica não ter melhorado o desempenho, ainda podemos ver, nos gráficos apresentados até aqui, que conseguimos resultados muito melhores que os melhores *F* do piso aleatório.

Gráfico 7: Resultados encontrados para todas as condições experimentais das três simulações.

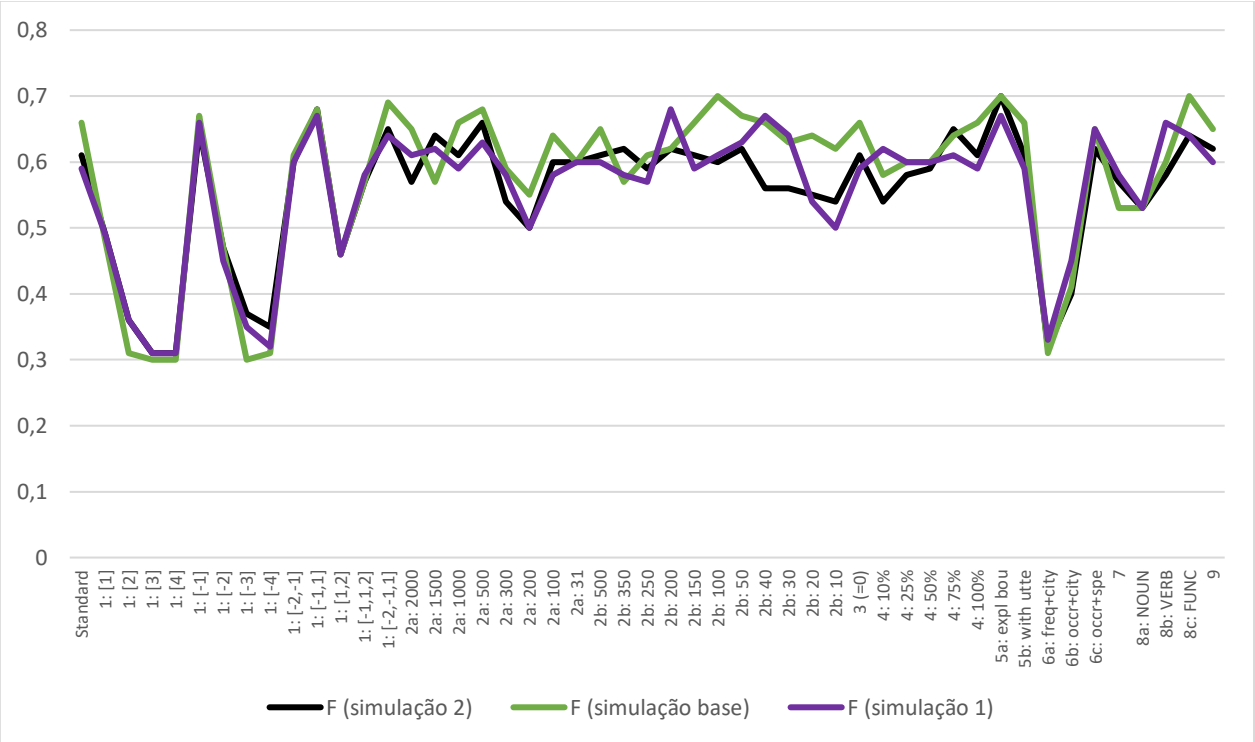


Tabela 17: Melhores resultados encontrados na *simulação base*, *simulação 1* e *simulação 2*.

Simulação base		Simulação 1		Simulação 2	
Experimento	**F**	**Experimento**	**F**	**Experimento**	**F**
Standard	0,66	Standard	0,59	Standard	0,61
1:[-2,-1,1]	0,69	1: [-1,1]	0,67	1: [-1,1]	0,68
2b: 100	0,7	2b: 200	0,68	2a: 500	0,66
3:noun	0,72	3:noun	0,72	3:	0,66
4: 100%	0,66	4: 10%	0,62	4: 75%	0,65
5a: expl. bou	0,7	5a: expl bou	0,67	5a: expl bou	0,7
6c: occr + spe	0,65	6c: occr+spe	0,65	6c: occr+spe	0,62
7	0,53	7	0,58	7	0,57
8c: FUNC	0,7	8b: VERB	0,66	8c: FUNC	0,64
9	0,65	9	0,6	9	0,62

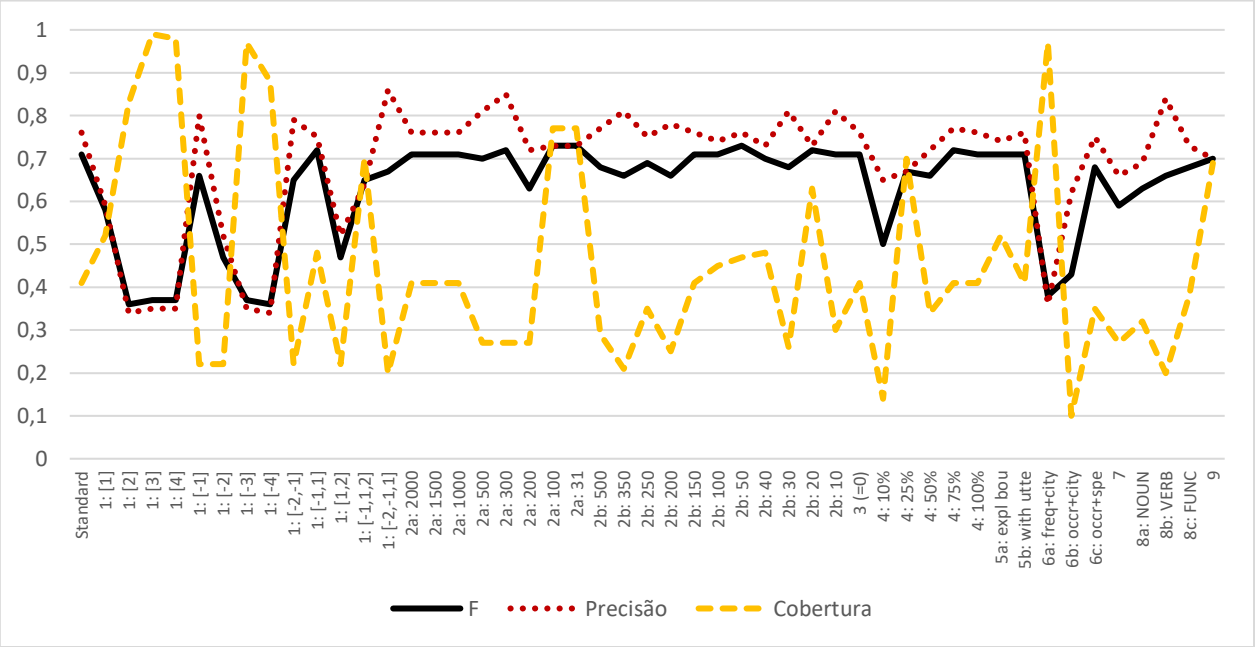
4.3 SIMULAÇÕES UTILIZANDO APENAS PALAVRAS COM INFORMAÇÃO MORFOLÓGICA

Diferente do que foi discutido na seção *4.2*, nesta seção demostramos o desempenho do modelo quando apenas as 650 palavras-alvo etiquetadas são contabilizadas para o cômputo da performance. Nessas simulações, podemos supor que o aprendiz faz a análise morfológica e aplica para todas as palavras-alvo que deseja categorizar. E é mais plausível esperar que o DAL (Dispositivo de Aquisição da Linguagem) da criança crie hipóteses morfológicas sobre todas as palavras. Essa é a motivação de avaliar o método usando apenas as palavras que receberam análise, isto é, uma forma de aproximar do que a criança estaria fazendo.

4.3.1 Simulação 3A: 650 palavras selecionadas

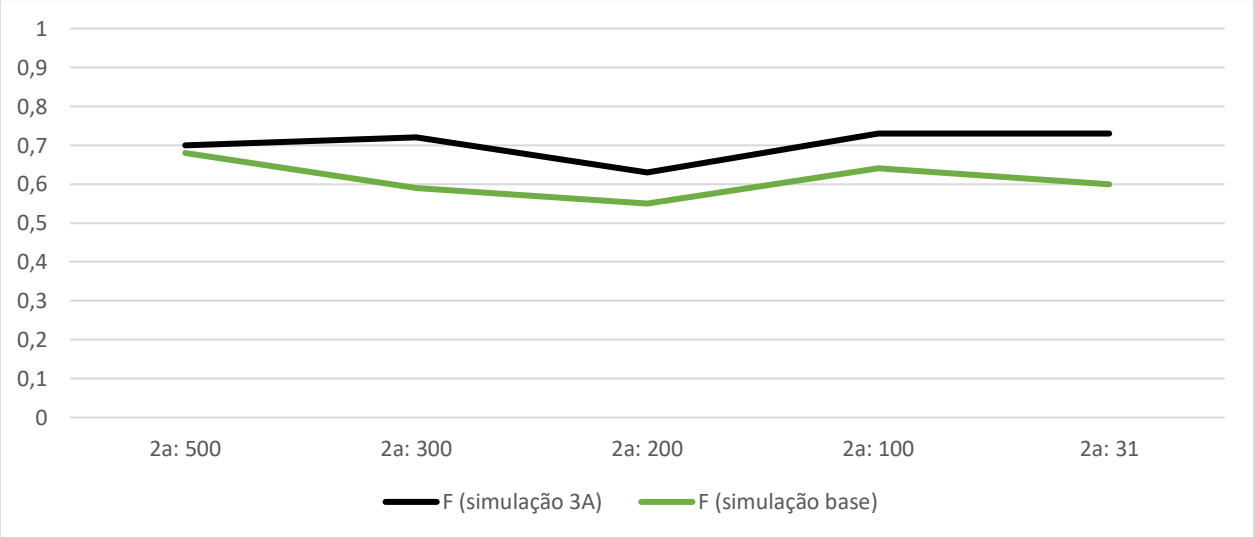
Durante a avaliação dos resultados, percebemos que a seleção das palavras alvo pode ser um fator que influencia a performance obtida. As simulações listadas até o momento utilizavam as palavras mais frequentes como o único critério para a seleção de palavras-alvo. Nessa nova simulação, restringimos a lista de palavras-alvo apenas àquelas com análise morfológica. Sendo isso, portanto, um critério de seleção complementar à frequência, podemos observar se a própria seleção de palavras influencia os resultados, mesmo quando a informação morfológica não é diretamente utilizada no processamento. O comportamento geral da simulação pode ser visto no *Gráfico 8*.

Gráfico 8: Resultados com F, precisão e cobertura para todas as condições experimentais da *simulação 3A*.



Analisando o *Gráfico 8*, constatamos certa tendência dos resultados com um *F* acima de 0,6. Porém, para que a comparação com a simulação base seja feita de uma forma mais justa, precisamos produzir resultados em que ambas as simulações estejam utilizando a mesma quantidade de palavras-alvo. Sendo assim, a comparação entre as condições experimentais pode ser vista no *Gráfico 9* para todas as condições em que temos a mesma quantidade de palavras nas duas simulações. Em todas as condições experimentais, temos que a simulação atual (*3A*) consegue alcançar valores F mais altos.

Gráfico 9: Resultados da simulação base e *simulação 3A* com as condições experimentais 2a utilizando a mesma quantidade de palavras-alvo.



Na *Tabela 18*, são apresentados os melhores resultados para ambas as condições. Não podemos fazer uma comparação direta entre cada uma das condições experimentais, pois as palavras são distintas. O objetivo aqui é mostrar que a forma de seleção das palavras-alvo pode ter impacto na performance: ocorre que as palavras com análise morfológica são mais facilmente categorizadas pelo modelo utilizando apenas informação distribucional. Como podemos ver, o modelo só não conseguiu o melhor desempenho na condição 8, onde a melhor variação deu o resultado de 0,68 e no modelo base (a) foi de 0,7.

Nossa conclusão é de que as condições experimentais com as palavras selecionadas pelos critérios que julgamos necessários, para receber a morfologia, se mostraram mais informativos do que os experimentos em que as palavras foram escolhidas pelo critério de seleção da frequência no *corpus*.

Tabela 18: Melhores *F* encontrados em cada uma das condições experimentais, na *simulação 3A* e *simulação base*, com a precisão e cobertura.

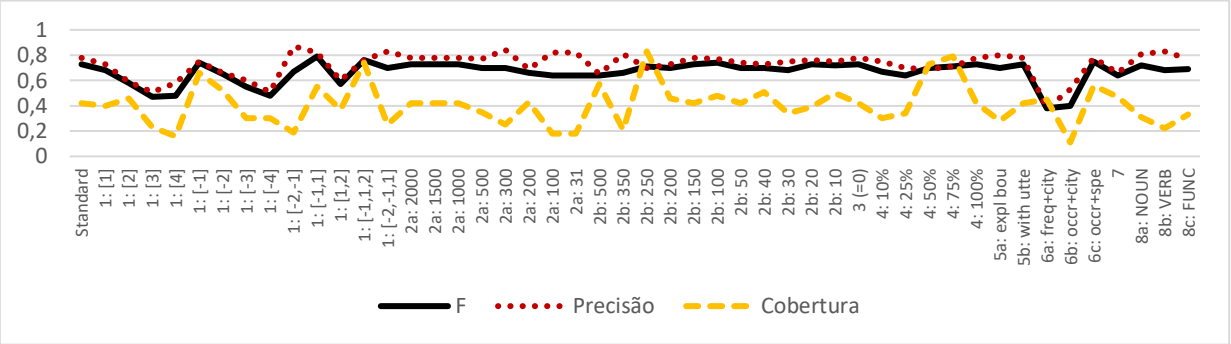
Simulação 3A				Simulação Base			
Experimento	**F**	**Precisão**	**Cobertura**	**Experimento**	**F**	**Precisão**	**Cobertura**
Standard	0,71	0,76	0,41	Standard	0,66	0,69	0,47
1:[-1,1]	0,72	0,75	0,48	1:[-2,-1,1]	0,69	0,74	0,41
2a: 100	0,73	0,73	0,77	2b: 100	0,7	0,74	0,44
3:verb	0,7	0,79	0,65	3:noun	0,72	0,72	0,67
4: 100%	0,71	0,76	0,41	4: 100%	0,66	0,69	0,47
5a: expl. bou	0,71	0,74	0,52	5a: expl. bou	0,7	0,73	0,51
6c: occr + spe	0,68	0,75	0,35	6c: occr + spe	0,65	0,74	0,26
7	0,59	0,66	0,27	7	0,53	0,54	0,5
8c: FUNC	0,68	0,73	0,38	8c: FUNC	0,7	0,73	0,45
9	0,7	0,7	0,69	9	0,65	0,65	0,72

4.3.2 Simulação 3B: 650 palavras com as informações morfológicas

Os resultados apresentados aqui foram obtidos a partir da utilização das 650 palavras etiquetadas como palavras-alvo com seus vetores morfológicos. No *Gráfico 10*, podemos observar os melhores *F*, a precisão e cobertura encontrados para todas as condições experimentais.

Observando o *Gráfico 10,* constatamos que os resultados também tenderam a se manter acima de 0,6, obtendo uma média de 0,67 em oposição aos 0,62 na *simulação base*. Diferente da *simulação 1*, os resultados do modelo apresentados aqui se mostraram melhores do que a *simulação base*. Conseguimos observar essa melhora no *Gráfico 11*, onde é possível ver o desempenho melhor na maioria dos experimentos. Entretanto, há um comportamento não esperado, para as condições experimentais 2a:100, 2a:31 e 2b:500, onde o modelo tem uma queda de desempenho, se comportando pior que a simulação 3A para essas mesmas condições.

Gráfico 10: Resultados com F, precisão e cobertura para todas as condições experimentais da *simulação 3B*.



Os melhores F para cada configuração podem ser vistos na *Tabela 19*. Podemos constatar que em todos os experimentos com a morfologia o modelo se comportou melhor, ao contrário do que apresentamos na seção *4.2*.

Gráfico 11: Resultados com os F encontrados para todas as condições experimentais nas *simulações 3B* e *3A*.

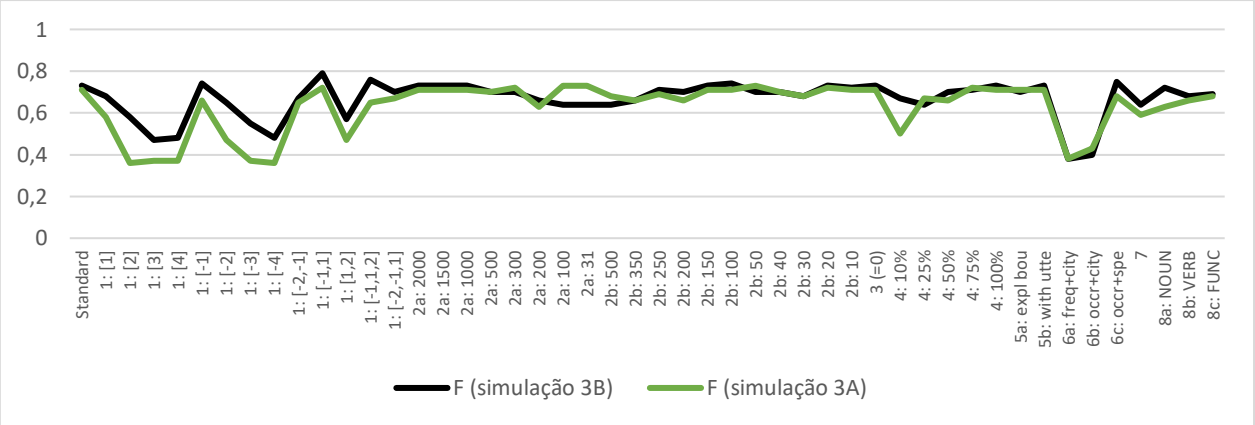


Tabela 19: Melhores F encontrados nas *simulações 3B* e *3A*.

Simulação 3B		Simulação 3A	
Experimento	**F**	**Experimento**	**F**
Standard	0,73	Standard	0,71
1:[-1,1]	0,79	1:[-1,1]	0,72
2b: 100	0,74	2a: 100	0,73
3:noun	0,75	3:verb	0,7
4: 100%	0,73	4: 100%	0,71
5a: expl. bou	0,73	5a: expl. bou	0,71
6c: occr + spe	0,75	6c: occr + spe	0,68
7	0,64	7	0,59
8a: NOUN	0,72	8c: FUNC	0,68
9	0,74	9	0,7

Desta vez, os resultados superiores obtidos usando a morfologia foram os esperados, já que ao inserir novas informações no modelo, nossa expectativa era de que o modelo tivesse melhor performance na tarefa de categorização, assim como entendemos acontecer no processo de

aquisição da linguagem. Ademais, observando a *simulação 3B*, percebemos que, com a morfologia, a janela mais informativa é a [-1,1], chegando a um F de 0,79 *versus* 0,72 na *simulação 3A* para a mesma janela. Esse resultado, considerando os parâmetros investigados, aponta para a conclusão de que o PB se organiza de forma que a informação mais local (morfologia e palavras adjacentes) é a que pode ser mais informativa sobre a categorização.

4.3.3 Simulação 3C: 650 palavras com informação morfológica e função

Dando continuidade aos experimentos, chegamos na última variação, em que utilizamos apenas as 650 palavras etiquetadas e os vetores morfológicos de cada uma com a função inserida. Dentro das configurações utilizadas até aqui, essa é a mais completa, em que todas as informações são inseridas. O resultado para cada condição experimental e suas variações pode ser visto no *Gráfico 12*.

Para essa simulação, não foi possível obter uma melhora de forma geral, com a média do *F* se mantendo em 0,67. No *Gráfico 13*, podemos ver como o modelo se comporta melhor em algumas condições experimentais e pior em outras. Ainda não é possível explicar o motivo de o modelo se comportar dessa forma, só o que podemos afirmar é que as funções deixaram o modelo menos preciso em alguns experimentos e mais preciso em outros. Em suma, não houve melhora (e nem piora) geral, quando nossa hipótese é a de que haveria melhora.

Gráfico 12: Resultado geral para todas as condições experimentais da *simulação 3C*.

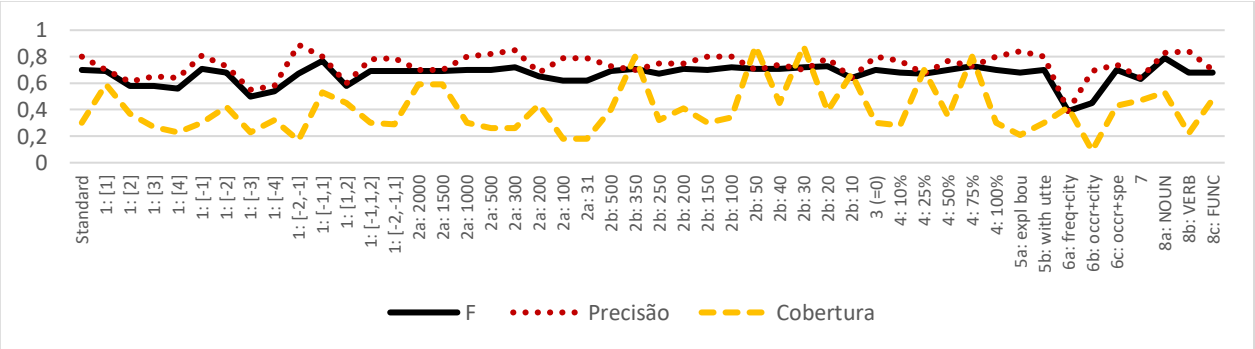
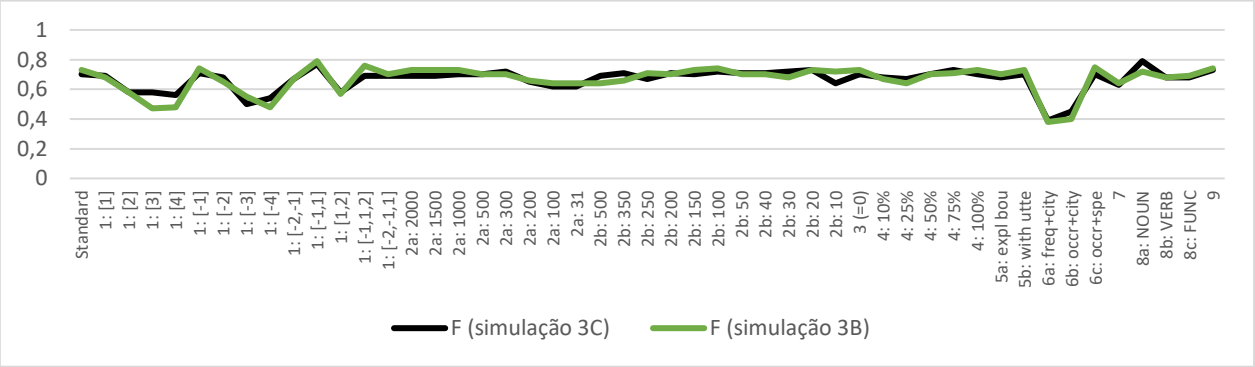


Gráfico 13: Melhores *F* para todas as condições experimentais das *simulações 3B* e *3C*.



A *Tabela 20* traz os melhores F para todas as 3 simulações que utilizam as informações morfológicas. A melhor simulação, se olharmos apenas para os melhores *F*, é a *simulação 3B*, que utiliza apenas a informação morfológica. A *simulação 3C* se comportou melhor nas condições experimentais que não são as mesmas da morfologia, tendo um desempenho melhor no experimento 8a: NOUN, onde todos os nomes são substituídos pela *etiqueta NOUN*. A partir dessa análise quantitativa, conseguimos perceber como determinadas mudanças ou configurações dos experimentos são mais ou menos informativas de forma geral.

Tabela 20: Melhores F das simulações que utilizam as 650 palavras etiquetadas como palavras-alvo.

Simulação 3B		Simulação 3C		Simulação 3A	
Experimento	**F**	**Experimento**	**F**	**Experimento**	**F**
Standard	0,73	Standard	0,7	Standard	0,71
1:[-1,1]	0,79	1:[-1,1]	0,77	1:[-1,1]	0,72
2b: 100	0,74	2b: 10	0,73	2a: 100	0,73
3:noun	0,75	3:noun	0,72	3:verb	0,7
4: 100%	0,73	4: 75%	0,73	4: 100%	0,71
5a: expl. bou	0,73	5b: with utte	0,7	5a: expl. bou	0,71
6c: occr + spe	0,75	6c: occr + spe	0,7	6c: occr + spe	0,68
7	0,64	7	0,63	7	0,59
8a: NOUN	0,72	8a: NOUN	0,79	8c: FUNC	0,68

Em um primeiro momento, minimamente, temos duas análises diretas que podemos fazer, a mais óbvia é que o modelo não utiliza de forma plausível a função e isso pode indicar um possível erro no modelo. A outra, se o modelo estiver correto, significa que a função pode ser considerada desnecessária no processo para o aprendiz, já que não há nem melhora, nem piora para o modelo.

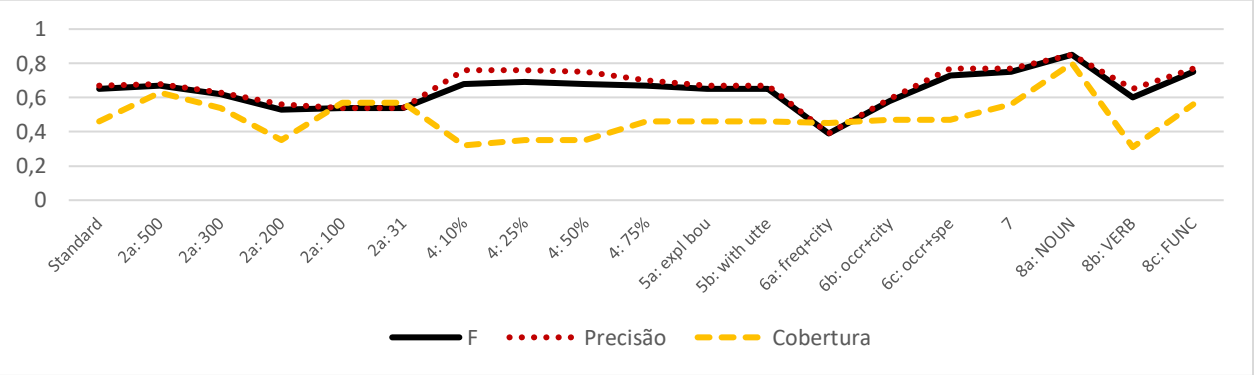
4.3.4 Simulação 4: utilizando as 650 palavras sem informação do contexto

Apesar de não ser uma simulação que possa ser comparada diretamente com as que foram apresentadas até o momento, como estamos investigando a informatividade da morfologia no

contexto da informação distribucional, é interessante investigar o quanto a morfologia pode ser útil se analisada de forma independente, sem qualquer análise distribucional do contexto das palavras. Para isso, também utilizamos apenas as 650 palavras selecionadas e o resultado geral pode ser visto no *Gráfico 14*. O fato de não levar em consideração o contexto como pista para o modelo faz com que as condições que variavam a janela possam ser representadas apenas pela condição *standard*, totalizando 19 condições experimentais.

Alcançando uma média do *F* de 0,64, a morfologia se mostrou bastante informativa, atingindo um valor médio maior do que a *simulação base*, que alcançou uma média *F* de 0,6, para as mesmas condições experimentais. Este é um resultado que parece interessante, pois sugere o quanto o aprendiz poderia aprender sobre as categorias, mesmo se descartasse todo o contexto das palavras e olhasse apenas para a morfologia. E, ainda, essa média encontrada se aproxima das médias *F* encontradas para as simulações que utilizam a morfologia com a informação distribucional, que é de 0,66 em cada uma, para as mesmas 19 condições experimentais.

Gráfico 14: Resultado da Simulação 4, sem análise distribucional do contexto, utilizando apenas o vetor morfológico.



4.4 ANÁLISE QUALITATIVA DA CONDIÇÃO STANDARD – SIMULAÇÃO 3B

Nesta seção, apresentamos a análise qualitativa da condição *standard*, da *simulação 3B*, que utiliza as 650 palavras que receberam informação morfológica como palavras-alvo e janela de contexto [-2,-,1,1,2]. Para essa configuração, o modelo encontrou 11 agrupamentos com corte no dendrograma de 0.63, atingindo um *F* de 0,73, precisão de 0,78 e completude de 0,42. Entre os 11 agrupamentos, apenas 4 tiveram grupos com 4 ou menos palavras, sendo 3 deles 100% puros e um contendo 2 pronomes e 1 verbo. Nos outros 7 foram agrupadas 40 ou mais palavras, sendo o maior grupo com 207 palavras. Esses agrupamentos podem ser vistos na *Tabela 21*, na qual observamos grupos mais heterogêneos (G4, G8 e G11), grupos mais homogêneos (G3, G5 e G7) e um agrupamento 100% preciso (G2). Apesar de encontrarmos apenas 4 palavras ou menos em 4

grupos, 3 deles são puros (G6, G1 e G9). O G2 é um grupo com 104 palavras e todas elas foram agrupadas corretamente. Para uma melhor visualização, trazemos as palavras encontradas em cada um dos seus respectivos conjuntos, sendo possível consultar o dendrograma na íntegra em: *Dendrograma para morfologia (link)*[20].

Tabela 21: Grupos extraídos do dendrograma da condição *standard - simulação 3B*.

	Grupos										
Categorias	G4	G3	G2	G8	G5	G7	G11	G6	G1	G10	G9
Adjetivo	10	2	-	27	11	-	7	-	-	-	-
Advérbio	2	-	-	1	3	-	4	-	-	-	-
Artigo	12	-	-	2	-	-	15	-	-	-	-
Nome	11	6	-	9	187	1	7	-	-	-	2
Preposição	2	-	-	-	-	-	-	-	-	-	-
Pronome	13	-	-	3	-	-	7	-	-	2	-
Verbo	4	109	104	30	6	40	4	4	2	1	-
Total	**54**	**117**	**104**	**72**	**207**	**41**	**44**	**4**	**2**	**3**	**2**

Olhando um pouco mais de perto para o grupo G2, *Tabela 22*, conseguimos verificar a existência de muitos verbos no infinitivo, o que mostra uma tendência, ou facilidade do modelo, para classificar essas palavras. Não podemos assumir que tal resultado se dá pelo compartilhamento da informação morfológica apenas, pois na simulação base, o modelo também classifica bem esta flexão verbal. Ainda, conseguimos ver verbos aparentemente flexionados, como em *lê, sai, cai* (em negrito) e verbos no passado, como é o caso de *caí* e *dormi*[21](em negrito). Então, apesar de aparentemente não ser um grupo estritamente homogêneo quanto à subclasse verbal, ainda assim o grupo se compõe apenas de formas verbais.

Tabela 22: Agrupamento G2 da condição *standard - simulação 3B*.

G2		Ocorrências
Verbo	fazer, ver, vê, brincar, contar, fala, ficar, pôr, ir, ser, tirar, pegar, falar, dizer, comer, pega, dar, buscar, levar, ter, chega, dormir, mostrar, come, escrever, tomar, guardar, sai, cai, botar, comprar, saber, passar, jogar, lê, andar, deixar, mexer, ouvir, vira, passa, cair, abrir, mexe, sentar, lavar, abre, leva, sair, brinca, conversar, ler, traz, cantar, desenhar, limpar, usa, joga, bate, entrar, chegar, chorar, cortar, estar, pintar, procurar, querer, trabalhar, trazer, gravar, ganhar, olhar, acabar, fechar, pedir, usar, chamar, bater, aperta, começar, perguntar, nadar, beber, vim, montar, achar, virar, saí, vir, descer, almoçar, jantar, dormi, caí, estragar, deitar, voltar, ligar, viajar, quebrar, puxar, esperar, trocar, ensinar	104
Total		**104**

[20] Link para o dendrograma completo: <https://bityli.com/cgPuM>.
[21] Devido a possíveis inconsistências das transcrições, não é possível afirmar que essas palavras não estavam no infinitivo.

Ainda analisando os grupos mais puros, vamos olhar para o G7, onde, a priori, temos um grupo heterogêneo, com 40 verbos e 1 nome. Entretanto, analisando a única palavra que foi categorizada como nome, *Tabela 23*, vemos que ela é homônima – o verbo *contar* na segunda pessoa do presente do indicativo (contas) e um substantivo feminino no plural (contas) – e o método que foi implementado ainda não trata esses casos. Sendo assim, podemos considerar esse grupo como mais um grupo com agrupamento ideal. Sobre os verbos que foram categorizados, conseguimos ver que nesse grupo temos apenas verbos flexionados, diferente do G2, onde a maioria são de verbos no infinitivo.

Tabela 23: Agrupamento G7 da condição *standard - simulação 3B*.

G7		Ocorrências
Verbo	estás, estão, estou, foste, gostas, foram, fizeste, viste, têm, for, conte, dormindo, temos, contasse, fizeram, dê, fazes, fosse, fazem, comendo, ouve, disseste, aprendeu, pondo, dando, estavas, machucou, chamava, conversando, procurando, ouvi, sentada, correndo, eram, lava, ficam, começou, estamos, gravando, tirando	40
Nome	contas	1
Total		**41**

Analisando os grupos que parecem ser menos homogêneos, mas ainda atentando-nos para aqueles com uma quantidade maior de verbos, temos o G3, *Tabela 24*. Nele, 8 palavras, das 117, foram categorizadas com outras categorias que não verbo, os adjetivos e os nomes. Todas as palavras consideradas como um erro pelo modelo, nesse grupo, são homônimas e como já dito, o modelo ainda não está preparado para lidar com palavras homônimas. Sendo assim, temos mais um grupo em que se poderia dizer que todas as palavras foram agrupadas corretamente. Com relação aos verbos, todos são flexionados, assim como no G7.

Tabela 24: Agrupamento G3 da condição *standard - simulação 3B*.

G3		Ocorrências
Verbo	olha, vai, tem, está, foi, vamos, quer, sabe, faz, pode, viu, deixa, vem, dá, põe, acho, chama, diz, fica, gosta, fez, tava, tinha, tô, tens, deu, espera, fazendo, mostra, caiu, anda, falou, tira, aconteceu, vão, senta, ficou, estava, tenho, disse, lembra, toma, acabou, vendo, acha, vi, deixo, contou, falando, canta, escuta, escreve, pegou, chegou, tirou, achou, saiu, quebrou, veio, comeu, brincando, ganhou, fecha, mora, passou, escreveu, começa, ponho, achei, faço, puxa, quebra, bateu, teve, deixou, olhando, levou, contando, procura, jogou, fiz, fico, pára, levanta, comprou, pede, estraga, tomou, falo, fazia, dormiu, chora, chorou, entra, esqueceu, ensinou, perdeu, entrou, desenha, pediu, abriu, deita, andando, brincou, explica, botou, acaba, desenhou, toca	109
Nome	conta, falta, tiro, pego, conversa, compra	6
Adjetivo	precisa, limpa	2
Total		**117**

O maior grupo encontrado, como era de se esperar, foi o grupo com maior concentração de nomes. Nesse grupo, temos um total de 207 palavras, *Tabela 25*, sendo que a maioria, 187, são nomes. As outras 20 palavras, se dividem em 11 adjetivos, 6 verbos e 3 advérbios. Apesar de ser um grupo heterogêneo, notamos uma forte tendência em agrupar os nomes, o que demonstra uma boa precisão, alcançando 90% nesse grupo (para nomes). Era de se esperar esse comportamento, visto que, assim como os verbos, é uma classe que representa uma quantidade significativa no *corpus*. Se olharmos para os grupos encontrados de forma separada, podemos chegar a uma precisão ainda maior, considerando que algumas palavras são homônimas ou, possivelmente, estão com a etiqueta no *corpus* de referência errada. Quanto ao grupo dos verbos, por exemplo, todas as palavras são homônimas, ainda que seja provavelmente mais rara a ocorrência de *papar* como nome, por exemplo. No que diz respeito ao grupo de adjetivos, temos a ocorrência de *professora*, que possivelmente está com uma etiqueta errada no *corpus* de referência, pois é mais comum o uso como nome. Uma possibilidade para o agrupamento desses elementos, além dos contextos similares em que eles aparecem, são os sufixos inseridos, já que não indicamos a função, mas apenas a existência, logo, os sufixos como *-inho, -o e -a* podem ter contribuído para que os grupos de nomes e de adjetivos ficassem juntos. Por fim, para os advérbios, a baixa taxa de ocorrências dessa categoria pode explicar o motivo da categorização.

Tabela 25: Agrupamento G5 da condição *standard - simulação 3B*.

	G5	Ocorrências
Nome	coisa, casa, história, mão, dia, pé, tia, carro, escola, historinha, vez, livro, hora, menina, pouquinho, coisas, boneca, cama, bola, estória, microfone, quarto, brinquedo, banho, aniversário, festa, menino, tempo, sapato, senhora, gato, maria, cara, porta, vestido, caixa, bicho, carrinho, dedo, desenho, juliana, cavalo, fita, jogo, música, casinha, anos, medo, cachorro, praia, perna, mesa, mundo, criança, calça, lobo, sala, cadeira, caneta, peixe, colo, negócio, barriga, horas, olho, pergunta, tio, galinha, folha, chupeta, volta, mala, porco, moça, rio, menininha, livrinho, urso, bolsa, senhor, brincadeira, coelho, remédio, pato, bolo, chucha, bolinha, piscina, força, gatinho, dinheiro, braço, ovo, calcinha, caixinha, filhinho, copo, brinquedos, tampa, escolinha, cola, ano, rosa, areia, fralda, branco, pessoas, bichinho, cozinha, amiga, pena, ginástica, vaca, prato, dica, amigo, letra, máquina, patinho, médico, janela, colégio, sacola, dias, pés, fio, olhos, estorinha, peça, passarinho, sopa, almoço, feira, página, aula, blusa, orelha, resto, gaveta, roda, peixinho, porquinho, palhaço, escova, plástico, ponta, girafa, sítio, pulseira, revista, número, quadro, livros, bicicleta, figura, prima, coelhinho, cachorrinho, crianças, peças, minutos, mamadeira, lenço, moço, pedaço, trabalho, problema, bichos, desenhos, buraco, pessoa, lua, espelho, boneco, laranja, cavalinho, graça, briga, rabo, pilha, teatro, beijinho, carta, cobra, dedinho, buraquinho, chuva	187
Adjetivo	novo, vermelho, baixo, meia, amarelo, outros, professora, pequenininho, tamanho, primeira, preto	11
Verbo	papar, bota, posto, colher, beijo, saia	6
Advérbio	pouco, todos, todas	3
Total		**207**

Dando sequência, vamos agora analisar os grupos mais heterogêneos, G4, G8 e G11. Entre esses, vamos verificar os grupos com mais palavras funcionais, que são o G4 e G11. Olhando para o agrupamento G4, na *Tabela 26*, apesar de termos uma tendência para agrupar as palavras funcionais, temos palavras divergentes da maioria agrupada: 11 nomes, 10 adjetivos, 4 verbos e 2 advérbios. Concentrando-nos no grupo dos nomes, concluímos que há uma tendência a nomes próprios, o que faz um certo sentido se olharmos para o agrupamento na *Tabela 25* onde a maioria dos nomes tem uma morfologia mais diversificada e no G4, temos uma agrupamento tendendo para os sufixos mais frequentes *-o*, *-a*. Vemos, ainda, algo similar nos adjetivos e, nos verbos, temos o agrupamento do verbo *ser*. A palavra *desse* ainda pode ser um erro de etiqueta no CTB, já que pode ser o pronome demonstrativo *desse*.

Tabela 26: Agrupamento G4 da condição *standard - simulação 3B*.

	G4	Ocorrências
Pronome	você, ela, ele, seu, qual, sua, teu, tua, dela, eles, vocês, dele, nossa	13
Artigo	um, na, uma, da, no, do, esse, essa, este, esta, aquele, aquela	12
Nome	filha, renata, filho, augusto, daniela, cor, daniel, luciano, filhinha, paulo, vezes	11
Adjetivo	bom, mesmo, outro, outra, grande, boa, linda, certo, nova, loiro	10
Verbo	sei, são, será, desse	4
Preposição	de, pra	2
Advérbio	muito, todo	2
Total		**54**

Fato similar ao que aconteceu no agrupamento G4 pode ser visto no agrupamento G11, *Tabela 27*, onde percebemos uma tendência para agrupar palavras funcionais, mas ainda assim formando um grupo heterogêneo que inclui 7 pronomes, 7 nomes, 7 adjetivos, 4 verbos e 4 advérbios. Entretanto, apesar de ser heterogêneo, existe certa tendência dentro dos subgrupos, onde temos *quantos*, *quantas*, *seus*, *suas* para os pronomes, por exemplo, ou ainda, *algum* e *algumas* para os advérbios (segundo a classificação de referência) e *pensei* e *pensava* para os verbos. O motivo para esse agrupamento ainda precisa ser investigado de forma mais acurada, analisando o contexto sintático em que essas palavras aparecem.

Tabela 27: Agrupamento G11 da condição *standard - simulação 3B*.

	G11	Ocorrências
Artigo	pro, dos, esses, essas, das, nessa, umas, nas, pelo, estes, pela, estas, aquelas, naquela, aqueles	15
Pronome	quantos, elas, teus, quantas, deles, suas, seus	7
Nome	mana, ajuda, meninos, causa, amigos, cores, mãos	7
Adjetivo	outras, cheia, cheio, grandes, duro, errado, pequenina	7
Advérbio	alguma, muita, algum, devagar	4
Verbo	gira, continua, pensei, pensava	4
Total		**44**

O último grupo a ser analisado, o G8, na *Tabela 28*, apesar de ser heterogêneo, contém dois grandes grupos, formados por verbos e adjetivos. Dos 30 verbos agrupados, temos um grupo, com

formas infinitivas e flexionadas, sendo as flexões referentes ao presente, passado e gerúndio, além das palavras *achas*, *gosto*, *desculpa*, *conto* e *giro*, no grupo dos nomes. Porém, provavelmente, *achas* está com a etiqueta errada e as demais palavra são homónimas - o verbo *gostar* na primeira pessoa do presente do indicativo (gosto) e um substantivo no masculino singular; o verbo *desculpar* na terceira pessoa do presente do indicativo (desculpa) e um nome feminino singular; o verbo *contar* na primeira pessoa do presente do indicativo (conto) e um nome masculino singular; o verbo *girar* na primeira pessoa do presente do indicativo (giro) e um nome no masculino singular - Nos 27 adjetivos encontrados, percebemos uma diversidade no sufixo também, mas, ainda assim, encontramos as combinações *coitada*, *coitadinha*, *coitadinho*. No grupo com 9 nomes, há casos em que poderíamos classificar como adjetivo, como em *velha*.

Tabela 28: Agrupamento G8 da condição *standard - simulação 3B*.

	G8	Ocorrências
Verbo	vou, quero, queres, vais, vá, queria, sabes, posso, gostou, sabia, deve, consegue, fui, conhece, falei, vês, acontecendo, coitado, quiser, dou, calhar, acontece, lembro, faltando, conheço, querida, conseguiu, feito, lembras, cansada	30
Adjetivo	bonito, lindo, bonita, claro, alto, sozinha, feio, direitinho, branca, bonitinho, coitada, gostoso, preciso, suja, mesma, sujo, pequenininha, coitadinha, pequena, pequeno, coitadinho, sozinho, feia, bonitinha, vermelha, bravo, brava	27
Nome	gosto, giro, dona, direito, achas, sono, desculpa, velha, conto	9
Pronome	quanto, nele, nela	3
Artigo	dessa, nesse	2
Advérbio	toda	1
	Total	**72**

Dos 7 grupos com 40 ou mais palavras apresentados, conseguimos homogeneidade em 3 deles, nos quais todas as palavras eram verbos, mostrando que o modelo com a morfologia tem uma certa facilidade de agrupamento dessa categoria sintática. Em outro grupo, conseguimos uma precisão de 90% para os nomes, sendo que algumas palavras eram homônimas e foram categorizadas com outras categorias sintáticas, o que faria com essa precisão fosse ainda maior se fossem consideradas como nomes. Nos 3 demais grupos, temos uma heterogeneidade alta, porém, ainda assim, o modelo demostrou que pode utilizar a informação distribucional para esse tipo de tarefa, ainda mais quando os subgrupos formados são analisados separadamente. Dessa forma, essa simulação aponta que a informação morfológica, mesmo sendo uma morfologia simplificada, contribuiu para uma melhora no modelo.

4.5 ANÁLISE QUALITATIVA DA CONDIÇÃO STANDARD – SIMULAÇÃO 3A

Nesta seção, apresentamos a análise qualitativa da condição *standard*, da *simulação 3A*, que é a *simulação base* restrita às 650 palavras selecionadas para receber as informações morfológicas, mas sem utilizar essa informação durante o processamento. Para essa condição, o modelo encontrou 8 agrupamentos com corte no dendrograma de 0,75, atingindo um *F* de 0,71, precisão de 0,76 e completude de 0,41. Entre os 8 agrupamentos, apenas 1 teve o grupo com 4 ou menos palavras. Nos outros 7, foram agrupadas 20 ou mais palavras, sendo o maior grupo com 200 palavras. Esses agrupamentos podem ser vistos na *Tabela 29*, na qual observamos grupos mais heterogêneos (G6, G3, G4 e G5), grupos mais homogêneos (G2, G1 e G8) e mais preciso, o G1, com 103 palavras, no qual todas foram agrupadas corretamente, de acordo com suas respectivas categoria. Para uma melhor visualização, trazemos as palavras encontradas em cada um dos conjuntos no dendrograma, em: *Dendrograma para morfologia (link)*[22].

Tabela 29: Grupos extraídos do dendrograma da condição *standard- simulação 3A*.

	Grupos							
Categorias	G6	G3	G2	G1	G8	G4	G5	G7
Adjetivo	27	7	3	-	1	13	6	-
Advérbio	7	1	-	-	-	1	1	-
Artigo	15	9	-	-	-	-	5	-
Nome	17	9	7	-	6	178	6	-
Preposição	2	-	-	-	-	-	-	-
Pronome	12	8	-	-	-	2	3	-
Verbo	7	3	115	103	59	6	7	4
Total	**87**	**37**	**125**	**103**	**66**	**200**	**28**	**4**

Analisando o G1, *Tabela 30*, é possível ver uma tendência para os verbos no infinitivo, assim como foi na *simulação 3B* e na *simulação base*. A *simulação 3B* que, assim como a *simulação base*, não utiliza as pistas morfológicas, demonstra a efetividade do método distribucional em agrupar verbos no infinitivo. Assim como observado na *simulação 3B*, há alguns verbos que não estão no infinitivo, como *vê*, *lê*, *caí*, *saí*, mostrando uma variedade nos tempos verbais, embora, como já dito, devido a possíveis inconsistências nas transcrições, não seja possível afirmar que esses verbos não sejam de fato formas infinitivas.

[22] Link para o dendrograma completo: <https://bityli.com/zOlUL>.

Tabela 30: Agrupamento G1 da condição *standard- simulação 3A*.

G1		Ocorrências
Verbo	fazer, ver, vê, brincar, contar, ficar, pôr, ir, ser, fazendo, tirar, pegar, falar, dizer, comer, dar, buscar, levar, vendo, ter, dormir, mostrar, escrever, tomar, guardar, falando, botar, comprar, saber, passar, jogar, lê, andar, deixar, mexer, ouvir, vira, cair, abrir, sentar, lavar, abre, leva, sair, brinca, conversar, ler, brincando, traz, cantar, desenhar, limpar, usa, joga, bate, entrar, chegar, chorar, cortar, estar, olhando, pintar, procurar, trabalhar, trazer, gravar, ganhar, olhar, acabar, fechar, pedir, ajudar, usar, chamar, bater, aperta, começar, perguntar, nadar, beber, vim, montar, achar, virar, saí, vir, descer, almoçar, jantar, dormi, andando, caí, estragar, deitar, voltar, ligar, viajar, quebrar, puxar, esperar, trocar, ensinar, aprender	103
Total		**103**

Dando sequência aos grupos mais puros, temos o G2, *Tabela 31*. São 115 verbos, representando um total de 92% de ocorrências. Porém, ao analisar as palavras que estão nos subgrupos Adjetivo e Nome, com exceção de *sozinho*, são homônimas. Como o modelo não trata dessas palavras, tais ocorrências são consideradas como erros. Porém, se nós levarmos em consideração esse conjunto de palavras, temos um grupo no qual 99% das ocorrências podem ser de verbo. Sobre os verbos que foram categorizados, a maioria é flexionada.

Tabela 31: Agrupamento G2 da condição *standard- simulação 3A*.

G2		Ocorrências
Adjetivo	precisa, limpa, sozinho	3
Nome	conta, falta, pergunta, tiro, pego, conversa, compra	7
Verbo	vai, tem, está, foi, vamos, quer, sabe, faz, pode, viu, deixa, vem, dá, põe, acho, chama, são, diz, fica, gosta, fez, tava, fala, tinha, tô, deu, espera, será, mostra, caiu, anda, falou, tira, aconteceu, vão, senta, ficou, pega, estava, tenho, disse, lembra, toma, acabou, chega, acha, come, vi, deixo, contou, sai, canta, cai, escuta, escreve, pegou, chegou, tirou, achou, saiu, passa, mexe, quebrou, veio, comeu, ganhou, fecha, mora, passou, escreveu, começa, ponho, achei, faço, puxa, quebra, bateu, dou, teve, deixou, levou, for, contando, procura, jogou, fiz, fico, pára, levanta, comprou, pede, estraga, tomou, falo, dê, fazia, dormiu, chora, chorou, entra, esqueceu, ensinou, fosse, perdeu, entrou, desenha, pediu, abriu, deita, brincou, explica, botou, acaba, desenhou, toca	115
Total		**125**

Ainda analisando os agrupamentos mais homogêneos, temos mais um em que o verbo é a maioria, o G8, que pode ser visto na *Tabela 32*, com uma situação similar ao que foi apresentada no agrupamento G2. Entretanto, nesse caso, todas as palavras podem ser verbos, então, temos um agrupamento com 100% de verbos, caso levemos em consideração a homonímia.

Tabela 32: Agrupamento G8 da condição *standard - simulação 3A*.

G8		Ocorrências
Adjetivo	preciso	1
Nome	gosto, ajuda, conto, passeio, tombo, contas	6
Verbo	vou, quero, queres, tens, vais, estás, queria, estão, estou, posso, gostou, sabia, deve, foste, consegue, fui, foram, fizeste, viste, conhece, falei, acontecendo, quiser, têm, querer, acontece, conte, dormindo, contasse, fizeram, lembro, faltando, conheço, fazes, fazem, comendo, ouve, disseste, aprendeu, pondo, dando, continua, conseguiu, machucou, chamava, conversando, procurando, ouvi, feito, sentada, lembras, correndo, eram, lava, ficam, começou, estamos, gravando, tirando	59
Total		**66**

Quanto aos grupos que parecem ser menos homogêneos, vamos iniciar com o G4 , *Tabela 33*, no qual há uma maior concentração de nomes, um total de 178 palavras, com um índice de 89% de participação no grupo. Assim como nos agrupamentos com o verbo, no agrupamento G4, temos situações de homonímias, *meia*, *professora*, *beijo*, dentre outras palavras, que poderiam ser classificadas como nome, mas como o modelo não trata dessas condições, são tidas como erros. Então, mesmo sendo um grupo com 5 categorias diferentes, ao levar em consideração esse fenômeno, temos um agrupamento com mais de 94% de nomes. Como possíveis "intrusos" no grupo, teríamos os pronomes e advérbios, que uma possível explicação seria a baixa quantidade de ocorrências.

Tabela 33: Agrupamento G4 da condição *standard - simulação 3A*.

G4		Ocorrências
Advérbio	pouco	1
Pronome	dela, dele	2
Verbo	papar, bota, posto, colher, beijo, saia	6
Adjetivo	grande, boa, linda, vermelho, baixo, meia, nova, amarelo, professora, pequenininho, tamanho, primeira, preto	13
Nome	coisa, casa, história, mão, dia, pé, tia, carro, escola, cor, historinha, vez, livro, hora, menina, pouquinho, coisas, boneca, cama, bola, estória, microfone, quarto, brinquedo, banho, aniversário, festa, menino, tempo, sapato, senhora, gato, maria, cara, porta, vestido, caixa, bicho, carrinho, dedo, desenho, juliana, cavalo, fita, jogo, música, casinha, anos, medo, cachorro, praia, perna, mesa, mundo, criança, calça, lobo, sala, cadeira, caneta, peixe, colo, negócio, barriga, horas, olho, tio, galinha, folha, chupeta, volta, mala, vezes, porco, moça, rio, menininha, livrinho, urso, bolsa, senhor, brincadeira, coelho, remédio, pato, bolo, chucha, bolinha, piscina, gatinho, dinheiro, braço, ovo, calcinha, caixinha, filhinho, copo, tampa, escolinha, cola, ano, rosa, areia, fralda, branco, pessoas, bichinho, cozinha, amiga, pena, ginástica, vaca, prato, amigo, letra, máquina, patinho, médico, janela, colégio, sacola, dias, fio, estorinha, peça, passarinho, sopa, almoço, feira, página, aula, blusa, orelha, resto, gaveta, roda, peixinho, porquinho, palhaço, escova, plástico, ponta, girafa, sítio, pulseira, revista, número, quadro, bicicleta, figura, prima, coelhinho, cachorrinho, crianças, minutos, mamadeira, lenço, moço, pedaço, trabalho, problema, buraco, pessoa, lua, espelho, boneco, laranja, cavalinho, graça, rabo, pilha, teatro, beijinho, carta, cobra, dedinho, buraquinho, chuva	178
Total		**200**

Nos grupos G3, G6 e G5, nenhuma categoria se mostrou dominante, porém, os padrões nos subgrupos dentro de cada agrupamento mostraram uma tendência de agrupamento. Por exemplo, no agrupamento G6, *Tabela 34*, temos um total de 31% de adjetivos, onde conseguimos observar os pares, *bonito*, *bonita*, e adjetivos com *-inho*, mostrando um determinado comportamento homogêneo. A mesma coisa podemos perceber no agrupamento de artigos, encontrando os pares *aquele*, *aquela*.

Tabela 34: Agrupamento G4 da condição *standard- simulação 3A*.

	G6	Ocorrências
Preposição	de, para	2
Verbo	vá, sabes, vês, coitado, calhar, querida, cansada	7
Advérbio	todo, alguma, todos, toda, muita, todas, algum	7
Pronome	qual, eles, vocês, quantos, elas, quanto, teus, quantas, nele, suas, nela, seus	12
Artigo	esta, aquele, pro, aquela, dos, esses, essas, das, dessa, nessa, nesse, umas, nas, pelo, aquelas	15
Nome	giro, dona, direito, achas, força, desculpa, meninos, brinquedos, pés, olhos, velha, livros, peças, bichos, desenhos, cores, briga	17
Adjetivo	bonito, lindo, bonita, claro, alto, sozinha, feio, direitinho, branca, outros, bonitinho, coitada, gostoso, outras, suja, mesma, sujo, pequenininha, coitadinha, pequena, pequeno, coitadinho, feia, bonitinha, vermelha, bravo, brava	27
Total		**87**

Outro agrupamento bastante heterogêneo, é o G3, *Tabela 35*, onde não temos um grupo dominante, sendo os dois maiores grupos, com um total de 9 ocorrências cada, nome e artigo. Para os nomes, conseguimos ver uma preferência para nomes próprios, já para os artigos conseguimos perceber o agrupamento de pares, *um*, *uma*, por exemplo. Para os pronomes também encontramos pares como *ela*, *ele*.

Tabela 35: Agrupamento G3 da condição *standard- simulação 3A*.

	G3	Ocorrências
Advérbio	muito	1
Verbo	olha, sei, desse	3
Adjetivo	bom, mesmo, outro, outra, novo, certo, loiro	7
Pronome	você, ela, ele, seu, sua, teu, tua, nossa	8
Artigo	um, na, uma, da, no, do, esse, essa, este	9
Nome	filha, renata, filho, augusto, daniela, daniel, luciano, filhinha, paulo	9
Total		**37**

Por último, temos o G5, *Tabela 36*, para o qual não conseguimos encontrar um padrão para os subgrupos. Nas 6 categorias levantadas, o verbo é a que mais ocorreu, entretanto, não é predominante, sendo contabilizadas 7 ocorrências. Ao contrário do que foi apresentado na análise qualitativa da *simulação 3B*, na qual os subgrupos tenderam a seguir algum padrão, nesse grupo não foi encontrado. No agrupamento G11, *Tabela 27*, da *simulação 3B*, temos o grupo mais heterogêneo, mas ainda conseguimos observar alguns padrões dentro dos subgrupos. E no grupo de artigos

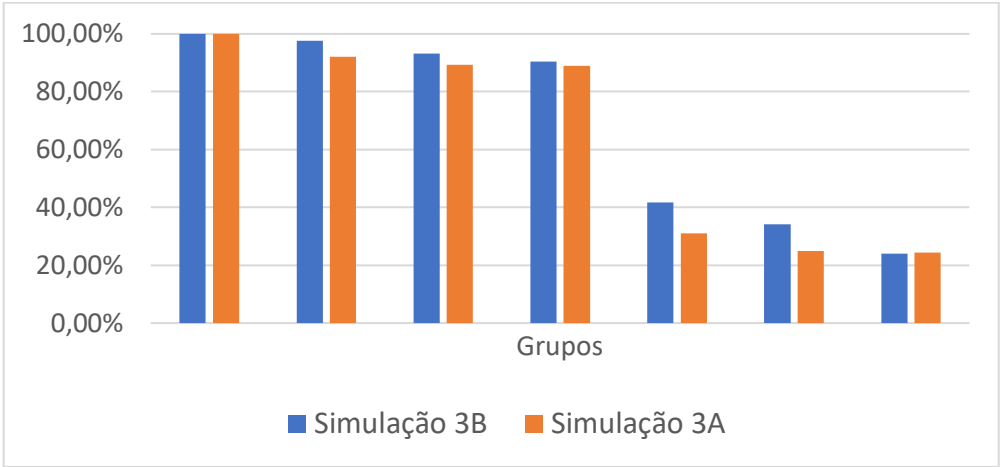
podemos ver um erro pela classificação do CTB, classificando *pela* como artigo e não como preposição.

Tabela 36: Agrupamento G5 da condição *standard- simulação 3A*.

	G5	Ocorrências
Advérbio	devagar	1
Pronome	quais, tuas, deles	3
Artigo	estes, pela, estas, naquela, aqueles	5
Nome	mana, sono, dica, causa, amigos, mãos	6
Adjetivo	cheia, cheio, grandes, duro, errado, pequenina	6
Verbo	gostas, temos, gira, estavas, pensei, dizes, pensava	7
	Total	**28**

Dos 7 grupos apresentados, observamos homogeneidade em 2 grupos, considerando a homonímia, para os verbos, e mais de 94% de ocorrências de uma classe em dois agrupamentos, um agrupamento com verbo e outro com nome. Se compararmos os resultados dos grupos encontrados nas duas simulações quanto à pureza dos grupos, sem levar em consideração as homonímias e olhando apenas para os grupos com mais de 20 ocorrências de palavras, conseguimos perceber uma pequena melhora na *simulação 3B*, conforme apresentado no *Gráfico 15*. A comparação é feita entre as categorias que obtiveram maiores ocorrências dentro do seu respectivo grupo.

Gráfico 15: Comparação entre as porcentagens de pureza das classes predominantes em cada grupo.



4.6 ANÁLISE QUALITATIVA DA CONDIÇÃO STANDARD – SIMULAÇÃO 4

Nesta seção, apresentamos a análise qualitativa da condição *standard*, da *simulação 4*, que utiliza apenas as pistas morfológicas, sem analisar o contexto. Para essa condição, o modelo

encontrou 8 agrupamentos com corte no dendrograma de 0.54, atingindo um *F* de 0,65, precisão de 0,68 e completude de 0,46. Entre os 8 agrupamentos, apenas 1, o G2, apresentou um conjunto com 8 palavras, no qual todas são verbos; nos outros 7, foram agrupadas 20 ou mais palavras, sendo que o maior grupo é composto por 267 palavras. Esses agrupamentos podem ser vistos na *Tabela 37*, em que observamos quais foram os grupos mais heterogêneos (G8, G5, G3, G7 e G6) e quais foram os mais homogêneo (G1 e G4), Para uma melhor visualização dos resultados, trazemos as palavras encontradas em cada um dos grupos em um dendrograma, disponível em: *Dendrograma sem contexto (link)*[23].

Tabela 37: Grupos extraídos do dendrograma da condição *standard- simulação 4*.

Categorias	G8	G5	G3	G4	G1	G7	G6	G2
Adjetivo	35	3	3	-	2	3	11	-
Advérbio	6	-	2	1	-	-	1	-
Artigo	10	-	17	-	-	2	-	-
Nome	23	30	8	-	3	89	70	-
Preposição	2	-	-	-	-	-	-	-
Pronome	-	12	9	-	-	1	3	-
Verbo	1	222	13	37	18	2	3	8
Total	**77**	**267**	**52**	**38**	**23**	**97**	**88**	**8**

Analisando o primeiro grupo mais homogêneo, o G4, *Tabela 38*, notamos que 97% das palavras são verbos e, assim como as outras simulações, em que houve o agrupamento de verbos no infinitivo, nesse agrupamento conseguimos ver essa tendência.. Provavelmente por conta do sufixo *-r*, foi acrescentado nesse grupo o advérbio *devagar*. Ainda se tratando dos verbos, aparecem outras flexões, como as formas no passado, *dormiu* e *aprendeu,* e o gerúndio *gravando*, que não compartilham do sufixo infinitivo das demais palavras.

Tabela 38: Agrupamento G4 da condição *standard- simulação 4*.

G4		Ocorrências
Advérbio	devagar	1
Verbo	foi, ser, será, buscar, dormir, guardar, papar, foste, ler, ganhou, foram, cortar, pintar, calhar, gravar, ganhar, for, dormindo, ajudar, dormiu, perguntar, ensinou, fosse, nadar, aprendeu, beber, montar, descer, almoçar, jantar, dormi, ligar, viajar, trocar, ensinar, aprender, gravando	37
Total		38

O segundo grupo mais homogêneo é o G1, no qual os verbos correspondem a 78% das ocorrências, seguidos por uma pequena quantidade de nomes e adjetivos, como é possível constatar na *Tabela 39*. Nesse agrupamento, notamos uma tendência para verbos no passado. As categorias

[23] Link para o dendrograma completo: <https://bityli.com/UdvsU>.

adjetivo e nome não compartilham os mesmos sufixos entre os verbos, porém, compartilham os sufixos *-inha*, *-inho* entre si, o que pode explicar o fato de estarem no mesmo agrupamento.

Tabela 39: Agrupamento G1 da condição *standard- simulação 4*.

G1		Ocorrências
Adjetivo	pequenininho, pequenininha	2
Nome	galinha, passarinho, minutos	3
Verbo	vou, viu, são, vi, fui, viste, dou, chorou, esqueceu, perdeu, pensei, machucou, saí, vir, correndo, eram, pensava, começou	18
Total		23

Quanto aos grupos mais heterogêneos, iniciamos a análise pelo G7, apresentado na *Tabela 40*. Constatamos que a categoria nome foi responsável por 91% das ocorrências. Nesse grupo, as palavras tendem a compartilhar o mesmo sufixo *-a* e o sufixo *-inha*. Devido ao algoritmo utilizado, há casos como o de *caixinha*, para a qual o sufixo escolhido foi o *-a*, compartilhando o mesmo sufixo dos demais nomes. Entretanto, existem casos como em *historinha*, onde o sufixo utilizado para a seleção foi o *-inha*.

Tabela 40: Agrupamento G7 da condição *standard- simulação 4*.

G7		Ocorrências
Pronome	nossa	1
Artigo	dessa, naquela	2
Verbo	continua, cansada	2
Adjetivo	meia, professora, primeira	3
Nome	coisa, casa, história, renata, escola, historinha, coisas, cama, estória, festa, maria, cara, porta, caixa, juliana, fita, música, casinha, praia, perna, mesa, dona, criança, calça, sala, cadeira, caneta, mana, barriga, pergunta, folha, chupeta, mala, ajuda, bolsa, chucha, piscina, força, desculpa, calcinha, caixinha, tampa, escolinha, rosa, areia, fralda, pessoas, cozinha, pena, ginástica, vaca, dica, letra, máquina, janela, sacola, estorinha, velha, peça, sopa, causa, feira, página, aula, blusa, orelha, gaveta, roda, escova, ponta, girafa, pulseira, revista, bicicleta, figura, prima, crianças, peças, mamadeira, problema, pessoa, lua, laranja, graça, briga, pilha, carta, cobra, chuva	89
Total		**97**

No G5,

Tabela **41**, os verbos apresentaram um índice de 83%. Para essa categoria, notamos uma variedade de sufixos, diante das diferentes flexões, o que é interessante, já que a única pista dada para o modelo foi o vetor morfológico. Isso pode indicar que a propriedade morfológica utilizada pelo modelo pode ser complementar ou redundante, de alguma forma, para o contexto, para essa situação. Quanto aos nomes, esses compartilham alguns sufixos com os verbos, o que pode indicar a razão para o agrupamento, assim como alguns pronomes, como *tuas*, *suas* (*-s*), que por mais que nesses casos o *-s* indique plural, já que não colocamos a função, o método não conseguiu distinguir entre esse morfema e o morfema indicativo de flexão, como em lembras (*-s*).

Tabela 41: Agrupamento G5 da condição *standard- simulação 4*.

	G5	Ocorrências
Adjetivo	grande, gostoso, grandes	3
Pronome	você, seu, qual, sua, teu, tua, vocês, teus, quais, tuas, suas, seus	12
Nome	mão, dia, pé, cor, vez, hora, microfone, gosto, giro, desenho, jogo, peixe, horas, olho, vezes, rio, brincadeira, achas, tiro, dias, pés, olhos, pego, peixinho, conversa, trabalho, desenhos, cores, compra, mãos	30
Verbo	olha, vai, tem, está, vamos, fazer, quer, sabe, faz, ver, pode, deixa, vem, vê, dá, põe, sei, acho, chama, diz, fica, brincar, gosta, fez, tava, fala, quero, tinha, queres, tô, ficar, tens, vais, deu, pôr, ir, estás, espera, vá, fazendo, tirar, pegar, mostra, caiu, anda, falar, falou, dizer, comer, tira, aconteceu, vão, queria, senta, ficou, sabes, pega, estava, tenho, dar, disse, estão, lembra, toma, acabou, estou, levar, vendo, ter, chega, mostrar, acha, desse, come, posso, escrever, tomar, deixo, falando, canta, cai, escuta, escreve, botar, gostou, comprar, saber, passar, jogar, lê, pegou, chegou, sabia, tirou, andar, deixar, mexer, deve, achou, ouvir, vira, passa, cair, abrir, mexe, sentar, quebrou, veio, lavar, comeu, abre, leva, brinca, gostas, consegue, conversar, brincando, traz, fecha, fizeste, cantar, mora, desenhar, conhece, passou, falei, escreveu, joga, bate, começa, vês, entrar, ponho, acontecendo, achei, chegar, faço, quiser, bota, quebra, bateu, têm, estar, teve, deixou, olhando, procurar, querer, trabalhar, podes, trazer, levou, acontece, olhar, procura, jogou, fiz, fico, acabar, pára, fechar, temos, levanta, pedir, fizeram, comprou, gira, pede, chamar, estraga, bater, tomou, lembro, falo, dê, fazia, aperta, conheço, fazes, entra, começar, querida, fazem, comendo, ouve, disseste, pondo, dando, querendo, colher, podia, quis, entrou, estavas, desenha, vim, conseguiu, pediu, abriu, achar, virar, chamava, conversando, procurando, ouvi, dizes, andando, caí, sentada, estragar, lembras, brincou, quebrar, explica, botou, lava, ficam, esperar, acaba, desenhou, estamos, tirando	222
Total		**267**

No G6, apresentado na *Tabela 42*, observamos uma tendência para o agrupamento de palavras com o sufixo *-o*. Além do sufixo *-o*, temos, também, o sufixo *-inho*, sendo compartilhado tanto na categoria nome quanto na categoria adjetivo. Os nomes representam 79% das ocorrências, sendo mais um dos casos em que temos uma quantidade maior de categorias, mas com um domínio expressivo por algum subgrupo.

O agrupamento G8,

Tabela 43, traz um agrupamento que tem adjetivos como categoria majoritária, com um índice de 45% das ocorrências. Nesse agrupamento, é possível observar uma variedade de sufixos (*-o*, *-inho*, *-a*), não havendo um sufixo padrão nas palavras agrupadas. E o agrupamento G3, *Tabela 44*, é o agrupamento mais heterogêneo. A categoria com a maior quantidade de ocorrências é o artigo, com um total de 32%, que se olharmos mais atentamente, identificamos uma tendência para o agrupamento dos sufixos *-s* e *-a*.

Tabela 42: Agrupamento G6 da condição *standard- simulação 4*.

	G6	Ocorrências
Advérbio	pouco	1
Pronome	quantos, quanto, quantas	3
Verbo	posto, beijo, feito	3
Adjetivo	certo, claro, baixo, alto, direitinho, amarelo, loiro, tamanho, duro, preto, errado	11
Nome	augusto, carro, livro, pouquinho, quarto, brinquedo, banho, aniversário, luciano, tempo, sapato, gato, vestido, bicho, carrinho, dedo, paulo, cavalo, anos, medo, cachorro, mundo, lobo, negócio, direito, porco, livrinho, urso, coelho, remédio, pato, sono, gatinho, dinheiro, braço, ovo, copo, brinquedos, ano, bichinho, prato, patinho, médico, colégio, fio, almoço, conto, resto, passeio, porquinho, palhaço, plástico, sítio, número, quadro, livros, coelhinho, cachorrinho, lenço, tombo, pedaço, bichos, buraco, espelho, cavalinho, rabo, teatro, beijinho, dedinho, buraquinho	70
Total		**88**

Tabela 43: Agrupamento G8 da condição *standard- simulação 4*.

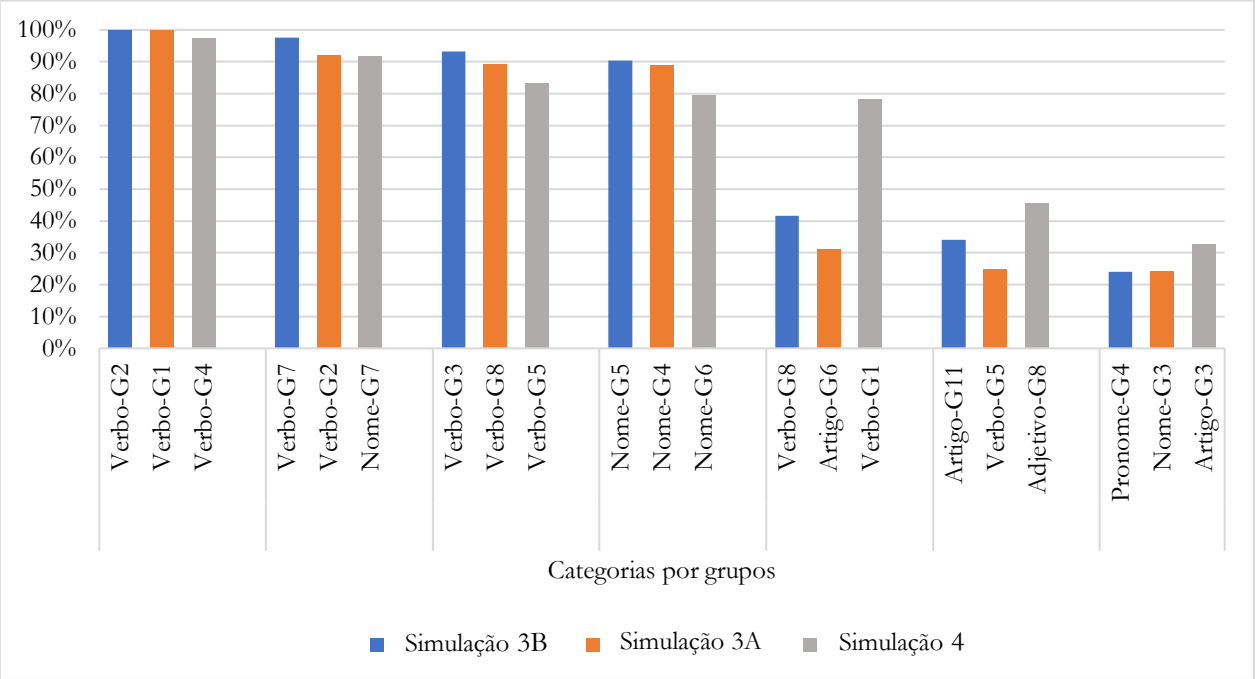
	G8	Ocorrências
Verbo	coitado	1
Preposição	de, pra	2
Advérbio	muito, todo, todos, toda, muita, todas	6
Artigo	na, da, no, do, pro, dos, das, nas, pelo, pela	10
Nome	filha, filho, tia, menina, boneca, bola, menino, filhinha, colo, tio, moça, menininha, bolo, bolinha, filhinho, meninos, cola, branco, amiga, amigo, moço, amigos, boneco	23
Adjetivo	mesmo, outro, outra, precisa, novo, bonito, lindo, linda, bonita, vermelho, sozinha, feio, nova, branca, outros, bonitinho, coitada, outras, preciso, suja, cheia, mesma, sujo, coitadinha, cheio, pequena, pequeno, coitadinho, sozinho, feia, bonitinha, vermelha, bravo, brava, pequenina	35
Total		**77**

Tabela 44: Agrupamento G3 da condição *standard- simulação 4*.

	G3	Ocorrência
Advérbio	alguma, algum	2
Adjetivo	bom, boa, limpa	3
Nome	conta, daniela, daniel, senhora, falta, volta, senhor, contas	8
Pronome	ela, ele, dela, eles, dele, elas, deles, nele, nela	9
Verbo	contar, contou, sai, saiu, sair, limpar, contando, conte, contasse, faltando, saia, voltar, toca	13
Artigo	um, uma, esse, essa, este, esta, aquele, aquela, esses, essas, nessa, nesse, umas, estes, estas, aquelas, aqueles	17
Total		52

Diferente das outras simulações, na *simulação 4* temos grupos mais heterogêneos, em que praticamente todas as classes se repetem pelos agrupamentos. Entretanto, como apresentado no *Gráfico 16*, o modelo tende a ter subgrupos com classes mais predominantes nos casos em que as outras simulações não tiveram um bom desempenho, G11 e G4, na Simulação 3B, e G5 e G3, na *simulação 3A*. Ou seja, utilizando apenas a morfologia, é possível alcançar um grau de pureza dentro dos grupos que se aproximam dos valores alcançados pelos grupos que utilizam contexto e morfologia, além de uma melhora nos grupos mais heterogêneos. E o fato de utilizarmos apenas a morfologia, sem contexto e sem a função, favorece o surgimento de tais grupos.

Gráfico 16: Comparação entre as porcentagens das classes predominantes em cada grupo.



Dessa forma, essa simulação aponta que a informação morfológica, mesmo sendo uma morfologia simplificada, contribui para uma melhora no modelo. Ainda mais interessante, pelo menos para o PB, a morfologia em si – ainda que simplificada, reiteramos – oferece bastante informação sobre as categorias das palavras, tornando a informação distribucional relativamente redundante. Como discutido nas conclusões, isso é ótimo para a criança, pois mostra que ela tem diferentes portas de entrada para as propriedades da língua, podendo recorrer a uma ou a outra em diferentes circunstâncias, a depender de sua saliência numa dada situação de exposição à língua.

4.7 TESTE DE SIGNIFICÂNCIA

Os dados mostram uma tendência de que a morfologia melhora o modelo, porém, é importante aplicar um teste estatístico que valide a nossa hipótese, de que ao inserir a morfologia a mudança é significativa. Para testar a hipótese de que a morfologia melhora o modelo, aplicamos o *teste T*, em que temos uma hipótese nula e uma hipótese alternativa. Aqui, assumimos a hipótese nula como aquela em que não existe diferença entre o modelo com morfologia e o modelo sem morfologia. Já a hipótese alternativa, a de que há diferença entre os modelos, é a nossa hipótese inicial. Assumindo o *teste T*, usei o teste bicaudal, pois não estamos procurando verificar se os dados são maiores ou menores, apenas queremos identificar se são diferentes a ponto de validar, ou não, a hipótese nula. E como temos as mesmas amostras antes e depois da morfologia, usamos o tipo

pareado. Por fim, assumimos um nível de significância (α) de 0,05, sendo bilateral, dividimos por 2 e temos o valor de 0,025, na prática, precisamos aplicar o *teste T* e encontrar um valor-p menor que 0,025 para negar a hipótese nula.

Aplicando o *teste T* entre os melhores F encontrados na *simulação base* e na *simulação 3A*, temos um valor-p de 0,000000000009567, ou seja, ele está abaixo do valor 0,025, então, podemos rejeitar a hipótese nula, de que os dois conjuntos são iguais, mostrando que existe uma significância ao escolher apenas as palavras que foram selecionadas para receber a morfologia. Conforme já observado, e agora validado, parece haver uma correlação positiva entre o grau de "transparência" morfológica das palavras em PB, isto é, a saliência de contrastes morfológicos, e a regularidade distribucional dessas palavras, de tal forma que o aprendiz se beneficia desta última.

Já era possível verificar visualmente a significância da diferença de performances entre a *simulação base* e *simulação 3A*. Porém, os valores encontrados entre as simulações *3A* e *3B* foram próximos um do outro, o que torna a aplicação do *teste T* ainda mais importante no caso destas duas simulações. O valor-p encontrado para essas amostras foi de 0,000206 e, assim como no resultado anterior, podemos rejeitar a hipótese nula, mostrando que a diferença entre os dois conjuntos de é estatisticamente significante. Isto significa que a informação morfológica de fato contribuiu para uma melhor performance do modelo distribucional. Finalmente, ao compararmos as simulações *3B* e *3C*, foi encontrado um valor-p de 0,881013, o que indica não haver diferença estatística entre as performances nestas duas simulações. Na *Tabela 45* é possível ver a descrição de cada simulação testada.

Por fim, contrastamos a *simulação 4*, que conta apenas com a informação morfológica, com as outras simulações, que utilizam o contexto e a morfologia. Comparando apenas as condições experimentais presentes na simulação 4 com as respectivas condições da *simulação 3A*, encontramos um valor-p de 0,969198, muito acima dos 0,025, não sendo possível rejeitar a hipótese nula. Um fato interessante acontece com essas condições experimentais, ao comparar as mesmas condições experimentais entre as simulações *3A* e *3B*, que anteriormente tinham rejeitado a hipótese nula, obtivemos um valor-p de 0,367173, validando a hipótese que antes tinha sido negada. Assim, observamos que para essas condições experimentais específicas (*standard*, 2a: 500, 2a: 300, 2a: 200, 2a: 100, 2a: 31, 4: 10%, 4: 25%, 4: 50%, 4: 75%, 5a: expl bou, 5b: with utte, 6a: freq.+city, 6c: occr+spe, 7, 8a: NOUN, 8b: VERB e 8c: FUNC), independe do que for utilizado, contexto, morfologia ou contexto com morfologia, não há como negar a hipótese nula. Todos os valores encontrados podem ser vistos na *Tabela 46*.

Tabela 45: Descrição das simulações utilizadas no teste de significância.

Simulação	Descrição
Simulação base	as nove condições experimentais implementadas na versão do modelo sem morfologia
Simulação 3A	simulação base restrita às 650 palavras selecionadas para receber as informações morfológicas, mas sem o uso dessas informações
Simulação 3B	650 palavras com as informações morfológicas
Simulação 3C	650 palavras com a informação morfológica e com as suas respectivas funções morfológicas
Simulação 4	650 palavras e usando apenas as informações da morfologia, sem utilizar a informação distribucional do contexto

Tabela 46: Resultados dos testes de significância aplicados.

Condição testada	Resultado
Simulação base e simulação 3A	0,000000000009567
Simulação 3A e simulação 3B	0,000206
Simulação 3B e simulação 3C	0,881013
Simulação 4 e simulação 3A	0,969198
Simulação 3A e simulação 3B – mesmos experimentos da condição 4	0,367173

Os resultados gerais apontam para a validade da hipótese de que a morfologia é impactante no modelo, mostrando que há um caminho a ser seguido, já que tanto o resultado comparando as simulações *base* quanto os resultados *3A* e *3B* tem um valor-p de 0,000206. Porém, ao fazer o recorte para os experimentos utilizados apenas na *simulação 4*, temos um resultado contrário, isso pode ser explicado pela seleção da amostra ser pequena, e ser uma amostra que não esteja representando o todo, já que não tem amostras de todos os tipos de experimento. Sendo assim, precisamos de mais resultados para uma amostra mais ampla para esse teste.

CONSIDERAÇÕES FINAIS

Ao observar os dados, é natural tendermos a uma análise do ponto de vista do adulto, do linguista, e tentarmos verificar o que é possível extrair das mais diversas pistas linguísticas; e, na perspectiva morfológica, tentarmos segmentar as palavras ao máximo para encontrar o que pode lançar luz à utilização da morfologia. Contudo, não necessariamente essa é a pista utilizada pela criança, afinal, pode ser que ela utilize apenas um dos *morfes* inicialmente e, aos poucos, vá entendendo a constituição morfológica das palavras.

Diante da natureza inicial desta investigação, algumas inquietações surgiram durante o andamento deste estudo: (i) *Qual seria o mínimo da morfologia que as crianças podem utilizar como pista para categorização?*; (ii) *Este mínimo seria similar à morfologia que foi considerada aqui?* Um dos caminhos para responder a essas perguntas é a utilização de modelos computacionais, como o que foi desenvolvido neste trabalho, combinados aos experimentos feitos com as crianças.

Os resultados encontrados neste trabalho se mostraram úteis para iniciar a tentativa de responder a tais perguntas. Entretanto, ainda não foi possível mensurar qual é o mínimo de morfologia que torna esta informação útil para a criança, mas conseguimos verificar que há melhora no modelo quando consideramos a morfologia como modelada neste trabalho. Além disso, os resultados da *Simulação 3A* mostraram que a tarefa pode começar na seleção do *input*, sendo esse um fator importante que, minimamente, ajuda na categorização de palavras, uma vez que apenas utilizando esse critério, conseguimos ter um aumento significativo no desempenho, como foi possível ver no teste de significância entre a *Simulação Base* e a *Simulação 3A*. Esse fator até então não estava sendo utilizado nos modelos, e precisa ser avaliado pois, aparentemente, as palavras selecionadas para receber morfologia são mais informativas distribucionalmente, o que poderia ajudar a criança, mesmo sem fazer uma análise morfológica. Ainda, podemos levar para outra perspectiva, mostrando que a morfologia ajuda na informação distribucional, sendo uma redundância para essa informação, auxiliando o aprendiz. Como verificamos nos testes de significância, não conseguimos negar a hipótese nula comparando os dados entre a *Simulação 3A* e a *Simulação 4*, então pode ser que essa redundância entre a informação distribucional e a morfologia seja verdadeira.

Dessa forma, como o objetivo principal era avaliar os dados e verificar como a morfologia se comportaria no modelo, partindo da hipótese de que ao inserir pistas morfológicas o modelo se comportaria melhor na tarefa de categorização, podemos concluir que os resultados deste trabalho mostraram que a morfologia tem um papel importante na aquisição das categorias, já que, seja com o contexto ou sozinha, foi um diferencial para o modelo, como foi possível ver nas análises e nos

testes de significância, tanto mais quando levamos em conta o caráter simplificado da morfologia no modelo. Com essa validação, é possível afirmar que o modelo está seguindo por um caminho coerente e que precisa de novas pistas linguísticas para se aproximar da categorização das crianças.

A investigação acerca da categorização durante o processo de aquisição da linguagem é uma tarefa árdua, com um longo caminho a ser percorrido: é necessário buscar respostas para perguntas já existentes e para outras que ainda serão feitas. Como por exemplo, ao alcançar melhores resultados com a utilização de uma morfologia simplificada, em comparação ao modelo sem a morfologia, o que aconteceria se o modelo utilizasse uma representação baseada nos manuais linguísticos? Conseguiríamos resultados melhores ou piores? A priori, poderíamos pensar que encontraríamos resultados melhores, porém, os resultados do *experimento 3C* mostraram que ao inserir a função morfológica, não foram obtidas melhoras significativas, constatação corroborada pelo teste de significância, evidenciando que não necessariamente ao inserir novas variáveis, conseguiríamos resultados melhores. Então, inserir uma análise morfológica exaustiva em um modelo distribucional pode trazer resultados contrários a hipótese de melhora do modelo. Além disso, para responder essa pergunta e as outras de forma mais consistente, o modelo precisa se aproximar de um aprendiz gradual, o que implica levar em consideração, além da morfologia, outros aspectos linguísticos.

Outra possibilidade é aplicar o mesmo método de seleção de palavras e morfologia para outras línguas e, depois, comparar com os resultados obtidos para o PB. Assim, conseguiríamos mensurar o quanto a morfologia é relevante para cada língua. Tomemos como exemplo o modelo desenvolvido por Redington *et al.* (1998), para o inglês, que foi utilizado como base para o trabalho apresentado aqui. Por fim, cabe destacar que o modelo não representa o aprendiz ideal, aquele aprendiz gradual, que consegue lidar com as mais diversas pistas linguísticas. Então, um próximo passo para o aprimoramento desse modelo é a inserção da gradualidade e analisar como ele se comporta utilizando as mesmas condições experimentais. Após isso, inserir pistas fonológicas, semânticas, prosódicas e sintáticas, sem ignorar nenhuma visão.

REFERÊNCIAS

ARONOFF, M., & FUDEMAN, K. (2011). What is morphology? (Vol. 8). John Wiley & Sons.

BULLINARIA, J. A., & LEVY, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior research methods, 39(3), 510-526.

BULLINARIA, J. A., & LEVY, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. Behavior research methods, 44(3), 890-907.

CÂMARA JUNIOR, J. M. (1970). Estrutura da língua portuguesa. Rio de Janeiro: Vozes.

CHRISTODOULOPOULOS, C., GOLDWATER, S., & STEEDMAN, M. (2010, October). Two decades of unsupervised POS induction: How far have we come?. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 575-584).

CLARK, A. (2003, April). Combining distributional and morphological information for part of speech induction. In 10th Conference of the European Chapter of the Association for Computational Linguistics.

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391-407.

DRESSLER, W. U., KILANI-SCHOCH, M., & KLAMPFER, S. (2011). How does a child detect morphology? Evidence from production. In Morphological structure in language processing (pp. 391-426). De Gruyter Mouton.

FARIA, P. e OHASHI, G. O. (2018). A aprendizagem distribucional no português brasileiro: um estudo computacional. Revista LinguíStica, 14(3): 128–156.

FARIA, P. (2019a). Aprendizagem de categorias de palavras por análise distribucional resultados adicionais para Português Brasileiro. Diacrítica, 33(2), 229-251. https://doi.org/10.21814/diacritica.415

FARIA, P. (2019b). The Role of Utterance Boundaries and Word Frequencies for Part-of-speech Learning in Brazilian Portuguese Through Distributional Analysis. In: Proceedings of the Workshop
on Cognitive Modeling and Computational Linguistics (NAACL'19), 152–159.

FARIA, P. (2020). Compreendendo a modelagem computacional de aquisição da linguagem. Veredas-Revista de Estudos Linguísticos, 24(1), 94-112.

FINCH, S., & CHATER, N. (1991). A hybrid approach to the automatic learning of linguistic categories. Artificial Intelligence and Simulated Behaviour Quarterly, 78, 16-24.

FINCH, S., & CHATER, N. (1992). Bootstrapping syntactic categories using statistical methods. Background and Experiments in Machine Learning of Natural Language, 229, 235.

HARRIS, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.

KAJIC, I., & ELIASMITH, C. (2018). Evaluating the psychological plausibility of word2vec and GloVe distributional semantic models. Technical Report CTN-TR-20180824-012. Centre for Theoretical Neuroscience, University of Waterloo. Waterloo, ON, Canada. doi: 10.13140/RG. 2.2. 25289.60004.

KILANI-SCHOCH, M., BALČIUNIENĖ, I., KORECKY-KRÖLL, K., LAAHA, S., & DRESSLER, W. U. (2009). On the role of pragmatics in child-directed speech for the acquisition of verb morphology. Journal of pragmatics, 41(2), 219-239.

LANDAUER, T. K., & DUMAIS, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2), 211.

MINTZ, T. H., NEWPORT, E. L., & BEVER, T. G. (1995). Distributional regularities of form class in speech to young children. In Proceedings-Nels (Vol. 25, pp. 43-54). University of Massachusetts.

MINTZ, T. H., NEWPORT, E. L., e BEVER, T. G. (2002). The distributional structure of grammatical categories in speech to young children. Cognitive Science, 26:393–424.

MINTZ, T. H., WANG, F. H., & Li, J. (2014). Word categorization from distributional information: Frames confer more than the sum of their (Bigram) parts. Cognitive psychology, 75, 1-27.

ONNIS, L., & CHRISTIANSEN, M. H. (2008). Lexical categories at the edge of the word. Cognitive Science, 32(1), 184-221.

PADO, S., & HOLE, D. (2019). Distributional Analysis of Polysemous Function Words. arXiv preprint arXiv:1907.10449.

PEARL, L. (2010). Using computational modeling in language acquisition research. Experimental methods in language acquisition research, v. 27, p. 163.

REDINGTON, M., CHATER, N., & FINCH, S. (1993). Distributional information and the acquisition of linguistic categories: A statistical approach. In Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society (pp. 848-853).

REDINGTON, M., CHATER, N., HUANG, C. R., CHANG, L. P., FINCH, S., & CHEN, K. J. (1995). The universality of simple distributional methods: Identifying syntactic categories in Mandarin Chinese. In Proceedings of the International Conference on Cognitive Science and Natural Language Processing. Dublin City University.

REDINGTON, M., CHATER, N., e FINCH, S. (1998) Distributional information: A powerful cue for acquiring syntactic categories. Cognitive science, v. 22, n. 4, p. 425-469.

RIBEIRO, M. G. C.. Morfologia da Língua Portuguesa. In: Aldrigue, A. C. de Sousa; Faria, E. M. Brito. (Org.). Linguagens: usos e reflexões. João Pessoa: UFPB/Editora Universitária, 2009, v. 3, p. 59-111.

MACWHINNEY, B. (1989). The CHILDES Project: Computational Tools for Analyzing Talk; Version 0.8. European Science Foundation.

MARGOTTI, F.; MARGOTTI, R. C., Morfologia do português, UFSC, UAB.— Florianópolis: LLV/CCE/UFSC, 2009.

SCHÜTZE, H. (1993, June). Part-of-speech induction from scratch. In 31st Annual Meeting of the Association for Computational Linguistics (pp. 251-258).

VALIAN, V. (1986). Syntactic categories in the speech of young children. Developmental Psychology, 22(4), 562–579. https://doi.org/10.1037/0012-1649.22.4.562