Volume 24, n.1 (2020)

ufjf

Veredas
Revista de Estudos Linguísticos
https://periodicos.ufjf.br/index.php/veredas/index

VEREDAS
revista de estudos linguísticos

# Compreendendo a modelagem computacional de aquisição da linguagem

# Understanding the computational modeling of language acquisition

Pablo Faria[1]

*Universidade Estadual de Campinas, Brasil*

RESUMO

É comum ouvir de aquisicionistas que modelos computacionais lhes são estranhos e relativamente inacessíveis. Porém, um frutífero intercâmbio de ideias e percepções mútuas é importante para o avanço de teorias de aquisição. Usando como exemplo o problema da aprendizagem distribucional de categorias sintáticas, apresentamos aqui um exemplo de processo que vai desde a descrição de um problema de aquisição, passando pela sua "tradução" para um problema de modelagem e chegando ao modelo propriamente dito. Neste percurso, tomamos consciência de como a modelagem impõe que o pesquisador explicite assunções, simplificações e decisões arbitrárias. Compreender o processo que conduz aos modelos computacionais nos tornará mais capazes, enquanto comunidade, de os avaliar criticamente e, mais que simplesmente apontar limitações, apontar também caminhos para seu aprimoramento ou para a plena exploração de seu potencial, através da elaboração de novos experimentos que respondam a novas perguntas.

PALAVRAS-CHAVE:

Aquisição da Linguagem, Modelagem Computacional, Aprendizagem Distribucional, Metodologia.

*Recebido em: 15/05/2020*
*Aceito em: 13/08/2020*

ABSTRACT

It is common to hear language acquisition researchers complain that computational models are too unfamiliar and relatively inaccessible. However, a fruitful exchange of ideas and mutual perceptions on models is important for the advance of acquisition theories. Taking as an example problem the distributional learning of syntactic categories, we present an example of a process which starts with a description of the acquisition problem, moves to its "translation" into a modeling problem, and ends with the model itself. In this route, we become aware of how modeling imposes that assumptions, simplifications, and arbitrary decisions, are made explicit. Understanding the process which leads us to computational models will enable us to, as a community, better evaluate them and, more than just pointing out limitations, also point out directions for further development, or for fully exploring their potential, through the formulation of novel experiments to answer novel questions.

KEYWORDS:

Language Acquisition, Computational Modeling, Distributional Learning, Methodology.

[1] E-mail: fariap@unicamp.br | ORCID: 0000-0002-4039-1769.

## 1. Introdução

Não é incomum ouvir de aquisicionistas que modelos computacionais lhes são por demais estranhos e inacessíveis a uma apreciação adequada. Tanto é assim, que o que se observa comumente nos dias de hoje são eventos científicos e publicações separadas para os estudos teóricos e experimentais de aquisição (p.e., o GALA, o ENAL, entre outros), de um lado, e estudos de modelagem computacional de outro (p.e., *workshops* CNCL e CogACLL). Por um lado, essa separação é compreensível, visto que aquisicionistas em geral têm formação mais profunda em linguística e em métodos experimentais, enquanto os "modeladores" têm formação computacional, sem negar aí as formações multidisciplinares, é claro. No entanto, essa separação é sem dúvida danosa e infeliz para nossa área como um todo. Na prática, ela acaba levando a uma espécie de indiferença tácita mútua, de modo geral, muito nociva.

É através do intercâmbio de ideias e percepções mútuas entre esses dois segmentos importantes da área que poderemos avançar na construção de teorias mais satisfatórias de aquisição. Como contribuição para viabilizar tais diálogos, portanto, este artigo traz um percurso possível na construção de um modelo computacional buscando, todavia, se esquivar da parafernália técnica computacional para focar no caráter mais substantivo desse processo.[2] Com isso, espera-se que o leitor relativamente alheio aos modelos computacionais possa contemplar e compreender melhor como e onde seu olhar de especialista, teórico e/ou experimental, pode incidir de modo a mais adequadamente avaliar e, assim, contribuir com tais modelos computacionais.

Com esse intento, o texto que segue está organizado da seguinte forma: na **seção 2**, é introduzida a noção de modelagem computacional, em que se busca pensar a noção de "modelo", além de encaminhar uma breve apreciação geral da história e do objetivo dessa abordagem. Na **seção 3**, é introduzido um problema de aquisição: o da aprendizagem distribucional de categorias sintáticas. É sobre esse problema que a construção de um modelo computacional é pensada na **seção** 4. Essa é a seção central do texto, em que a integração dos dois domínios, o da descrição linguística e o de sua "tradução" como um problema de modelagem é feita. A **seção 5** explora brevemente as possibilidades que um modelo abre, quanto a experimentações e manipulações de variáveis. Finalmente, a **seção 6** traz considerações finais em que refletimos sobre a relação entre

[2] Gostaria de agradecer aos comentários e observações valiosos feitos pelos pareceristas e que contribuíram para uma versão final mais clara do texto original que atende ainda mais aos seus objetivos.

modelagens e a realidade empírica.

**2. O que é modelagem computacional de aquisição?**

Construir modelos é buscar aproximações cada vez maiores, de forma gradual e sucessiva, de problemas do mundo real. Mas essas aproximações não podem ser tentativas ingênuas de reproduções da realidade, como nos mostra o conto de Borges (1974 *apud* BORGES NETO, 2004, p.69) sobre o rigor da ciência: no afã de criar o mapa perfeito, cartógrafos terminaram com um mapa que tinha o tamanho do império e coincidia pontualmente com ele. O resultado é um mapa inútil(izável). Um modelo, portanto, precisa capturar a estrutura e a dinâmica essenciais do fenômeno, de modo que, para fins de compreensão da realidade, permita manipular e observar a interação de todas as variáveis potencialmente envolvidas no problema. Em outras palavras, o modelo, assim como um mapa, nem pode ser demasiado simplório – omitindo informações importantes, como "limites", "obstáculos" e "marcos importantes" do "território" – nem pode ser demasiado detalhado, ao ponto mesmo de impedir uma apreensão global do mesmo. Idealmente, a uma distância razoável e com base no instrumental adequado, um bom modelo permite ao nosso olhar apreciar o funcionamento do todo e também o de suas partes.

É com essa finalidade que modelos computacionais de aquisição da linguagem são criados e é com base nela que são também avaliados. Tais modelos começaram a aparecer na década de 1970. Em Pinker (1979), vemos uma tentativa pioneira de avaliar o panorama das modelagens computacionais à época, na forma de simulações cognitivas e de inteligência artificial, além de ainda avaliar modelos matemáticos (teorias de aprendibilidade) e teóricos (modelos transformacionais). Estudos computacionais seguiram num crescente desde então, tanto em abordagens simbólicas, quanto com a retomada dos modelos conexionistas na década de 1980, entre outras abordagens. O progresso na área não se dá, no entanto, como uma linha evolutiva em que modelos mais elaborados e completos substituem os anteriores: sua evolução se dá mais pela variedade de abordagens e de questões que se busca responder e aí está sua pujança. Entre os esforços para apresentar panoramas mais atualizados da área, indico os levantamentos de Seidenberg (1997), Kaplan et al. (2008), Frank (2011) e Yang (2011). Uma ampla e criteriosa revisão de modelagens do problema da aquisição lexical pode ser vista também em Beraldo (2020).

Modelagens nos ajudam primordialmente a compreender *como se dá a aprendizagem*, isto

é, os mecanismos e condições de aprendizagem (ver PEARL, 2010, para uma introdução bastante didática a modelagens computacionais em aquisição). Mas elas também lançam luz sobre as duas outras questões, visto que estudos de aquisição envolvem três questões basilares:

a. *O que é aprendido?*
b. *De que modo essa aprendizagem ocorre no tempo?*
c. *Como essa aprendizagem se dá?*

A questão (a) envolve explicitar aquilo que se supõe ser o *alvo* da aquisição, ou seja, é a especificação do conhecimento adquirido. Nesse sentido, são as teorias linguísticas que nos fornecerão caminhos para essa especificação. Em (b), o objetivo é descrever aspectos desenvolvimentais da aquisição da linguagem, identificando marcos que envolvam mudanças importantes (p.e., a emergência das primeiras palavras), variações no ritmo de aprendizagem de um dado aspecto (p.e., a "explosão vocabular" no final do segundo ano de vida), os pontos em que se diz que algo foi adquirido, entre outras coisas. Finalmente, responder à questão (c) significa especificar os mecanismos e as condições de aprendizagem para os fenômenos investigados. Por mecanismos, entende-se estratégias, heurísticas, operações de generalização e analogia, e quaisquer outros meios e recursos através dos quais a criança transforma os dados da experiência em conhecimento. Dados da experiência, por sua vez, fazem parte das *condições* de aprendizagem: o que a criança tem como "dado de entrada" para a aquisição? Mas, além disso, pode-se investigar se e quais outros aspectos da experiência estão envolvidos na aprendizagem como, por exemplo, o papel da atenção conjunta (TOMASELLO, 2009).

Desse modo, a modelagem computacional propicia unir linguística, ciência da computação e psicologia do desenvolvimento, podendo ser utilizada para examinar uma ampla gama de questões sobre o processo de aquisição de linguagem. Uma vez que o objetivo de um modelo é ser uma simulação do problema, ele permite fazer manipulações e experimentações que seriam difíceis – senão impraticáveis – de realizar com crianças (temos exemplos concretos na seção 5). Assim, desde que os modelos computacionais levem em consideração as teorias linguísticas e psicolinguísticas, no intuito de serem simulações plausíveis da cognição (PINKER, 1979; PEARL, 2010), seremos capazes de obter resultados interessantes para a formulação de novas hipóteses sobre a aquisição de linguagem.

**3. O que é aprendizagem distribucional de categorias sintáticas?**

Entre os vários desafios na aquisição da linguagem, a criança precisa categorizar sintaticamente[3] as palavras da língua (VALIAN, 1986; GOODLUCK, 1991; GUASTI, 2002; entre outros), visto que essas assumem papéis bem definidos nas sentenças. Em muitas línguas, a existência de categorias sintáticas se reflete em certas regularidades distribucionais. Por exemplo, em uma oração que comece com "Ele", em português, é bastante natural esperar um verbo vindo logo a seguir, assim como se espera um sintagma nominal ou uma oração encabeçada por verbo infinitivo não flexionado depois da preposição em "Ele gosta de". Vários estudos mostram a sensibilidade de crianças a essa informação e que estas a usam ativamente para classificar sintaticamente palavras novas introduzidas na fala dirigida a elas (LANDAU & GLEITMAN, 1985; NAIGLES, 1990; BERNAL ET AL., 2007, entre outros). Portanto, a existência de tais regularidades distribucionais nos enunciados faz com que a evidência distribucional seja também uma fonte de informação potencialmente importante no processo de aquisição de categorias sintáticas, como enfatizam Redington et al. (1998).

Segundo Harris (1954, *apud* REDINGTON ET AL., 1998), a categoria de uma palavra pode ser descrita como a "soma" dos contextos em que ela ocorre. Em outras palavras, os tipos de contextos em que uma palavra ocorre podem ser usados como evidência de seu perfil gramatical, permitindo, assim, compará-la a outras palavras e, a partir disso, dividir o léxico em categorias. Vejamos como essa abordagem pode funcionar na prática. O que precisamos, basicamente, é de um corpus de fala (ou escrita) para que possamos listar as palavras a categorizar e levantar seus contextos de ocorrência. Vamos tomar algumas das palavras mais frequentes da fala dirigida à criança nos corpora de aquisição "Projeto Aquisição da Linguagem Oral" (CEDAE/UNICAMP[4]) e os de português disponíveis na base CHILDES[5], utilizado em Faria e Ohashi (2018) e Faria (2019a, 2019b): *que, é, de, com* e *vai*. Como contexto relevante neste exemplo, tomemos: $w_{i-1}$ $w_i$ $w_{i+1}$ (as palavras imediatamente anterior e posterior a $w_i$). Ao buscarmos no corpus utilizado no modelo os

[3] Na literatura sobre o assunto, as expressões "syntactic category" e similares são amplamente usadas e parte de uma já longa tradição, vide o texto de Valian (1986), por exemplo. É fato que a morfologia também inclui o conceito de categorias ou classes de palavras, tanto que muitas vezes se fala em categorização morfossintática, para abranger as situações em que há efeitos dos dois níveis linguísticos determinando o comportamento das palavras. Assim, optei por adotar a designação "categoria sintática" neste texto, especialmente por coerência com o tipo de aprendizagem em foco: ao olhar para regularidades distribucionais, a ordem se torna o fator (sintático) decisivo.

[4] Acessível em http://eulalio.iel.unicamp.br/sys/audio/albums.php?action=show&album=18. Último acesso em 24/08/2020.

[5] Acessível em https://childes.talkbank.org/data/Romance/Portuguese/. Último acesso em 25/08/2020.

10 primeiros enunciados encontrados para cada uma destas palavras, extraindo a janela distribucional de cada um, encontramos o seguinte[6]:

Tabela 1 – Os 10 primeiros contextos encontrados para palavras frequentes no corpus. O "." indica começo ou final de sentença.

| ***que*** | ***é*** | ***vai*** | ***De*** | ***com*** |
|---|---|---|---|---|
| o_que_aconteceu | ._é_. | você_vai_menina | brinquedo_de_armar | dia_com_o |
| o_que_você | água_é_muito | ._vai_esconder | armar_de_montar | ficou_com_medo |
| o_que_. | ._é_. | não_vai_passar | é_de_por | brincar_com_o |
| o_que_tá | que_é_que | que_vai_tirar | é_de_olhar | tá_com_fome |
| por_que_. | que_é_que | ._vai_achar | mais_de_botinha | ._com_açúcar |
| ah_que_que | não_é_assim | ._vai_por | ou_de_sapatinho | lá_com_a |
| que_que_é | não_é_assim | cê_vai_fazer | hora_de_dançar | tá_com_preguiça |
| é_que_tem | como_é_que | caixinha_vai_. | gosta_de_balinha | ._com_medo |
| ._que_é | que_é_que | não_vai_traze | gostou_de_ir | brincar_com_chinelo |
| é_que_tem | quem_é_que | ._vai_dançar | gostou_de_ir | brincar_com_ele |

Nos atendo tão somente à pequena amostra da Tabela 1, podemos notar algumas recorrências de contextos inteiros e outras tantas parciais (mesmo item anterior ou posterior). Por exemplo, "que" parece ser frequentemente precedido por "o" (4 vezes), ao passo que "é" é 3 vezes precedido pelo "que". O verbo "vai" costuma iniciar enunciados (4 vezes), enquanto a preposição "de" recorre com o verbo "gostou" (2 vezes) e sucede o verbo "é" (2 vezes). Finalmente, a preposição "com" sucede 3 vezes o verbo "brincar". Esse tipo de informação, ajuda a perceber a diferença entre os itens, porém precisamos tanto distinguir como também agrupar itens por semelhança categorial. Assim, outras informações da tabela vão contribuir para que possa ver coisas em comum entre os itens.

Por exemplo, ambos os verbos "é" e "vai" iniciam sentenças, como também sucedem elementos como "não" e "que" na tabela acima. Porém, a mesma tabela já não permite encontrar coisas em comum entre as preposições "de" e "com" (que, como falantes da língua, sabemos

[6] Estamos deixando de lado aqui várias questões relevantes, mas que não tocam o cerne do que queremos discutir aqui. Entre elas, como tratar alternâncias como "você/cê" e "está/tá", o tratamento de transcrições "como se fala", como em "trazê", e a diferença de sonoridade entre os dois "que" em "O que que...". Do ponto de vista da aquisição, trata-se de questões centrais, visto que a criança recebe exatamente esse *input* oral durante sua fase de aquisição e desenvolvimento da linguagem.

serem preposições). Isso mostra que muito mais dados seriam necessários para que a informação distribucional possa ser de fato útil para capturar relações categoriais entre palavras. Ademais, apenas *atestar* a co-ocorrência de uma palavra-alvo com um dado item lexical pode não bastar para categorizá-la eficientemente. Talvez seja necessário usar também a *frequência* dessas co-ocorrências. Ainda assim, são necessárias distinções adicionais. Parece bastante sensato supor, por exemplo, que ser precedido por um dado item não é o mesmo que ser sucedido por ele. Por exemplo, "que" é precedido 2 vezes por "é", que também o sucede outras duas vezes. Seria o caso de contar quatro co-ocorrências de "que" e "é" simplesmente? Não parece ser o ideal, pois obscurece possíveis aspectos sintáticos, já que a ordem é relevante nas línguas naturais. Se contabilizarmos separadamente e, ademais, unificarmos os dados de todas as palavras-alvo, poderíamos produzir uma tabela como a Tabela 2, chamada de *matriz de co-ocorrência*:

Tabela 2 – Matriz de co-ocorrência entre palavras-alvo e itens contextuais. Os itens <I> e <F> funcionam como marcadores explícitos de início e fim de enunciado, respectivamente.

| $w_{i-1}$ | *que* | *é* | *vai* | *de* | *com* | $w_{i+1}$ | *que* | *é* | *vai* | *de* | *com* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| o | 4 | | | | | <F> | 2 | 2 | 1 | | |
| é | 2 | | | 2 | | é | 2 | | | | |
| por | 1 | | | | | tem | 2 | | | | |
| ah | 1 | | | | | aconteceu | 1 | | | | |
| que | 1 | 3 | 1 | | | você | 1 | | | | |
| <I> | 1 | 2 | 4 | | 2 | tá | 1 | | | | |
| não | | 2 | 2 | | | que | 1 | 5 | | | |
| água | | 1 | | | | assim | | 2 | | | |
| como | | 1 | | | | muito | | 1 | | | |
| quem | | 1 | | | | menina | | | 1 | | |
| você | | | 1 | | | esconder | | | 1 | | |
| cê | | | 1 | | | passar | | | 1 | | |
| caixinha | | | 1 | | | tirar | | | 1 | | |
| gostou | | | | 2 | | achar | | | 1 | | |
| brinquedo | | | | 1 | | por | | | 1 | 1 | |
| armar | | | | 1 | | fazer | | | 1 | | |
| mais | | | | 1 | | trazê | | | 1 | | |
| ou | | | | 1 | | dançar | | | 1 | 1 | |
| hora | | | | 1 | | ir | | | | 2 | |
| gosta | | | | 1 | | armar | | | | 1 | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| brincar | | | | | 3 | montar | | | | 1 | |
| tá | | | | | 2 | olhar | | | | 1 | |
| dia | | | | | 1 | botinha | | | | 1 | |
| ficou | | | | | 1 | sapatinho | | | | 1 | |
| lá | | | | | 1 | balinha | | | | 1 | |
| | | | | | | o | | | | | 2 |
| | | | | | | medo | | | | | 2 |
| | | | | | | fome | | | | | 1 |
| | | | | | | açúcar | | | | | 1 |
| | | | | | | a | | | | | 1 |
| | | | | | | preguiça | | | | | 1 |
| | | | | | | chinelo | | | | | 1 |
| | | | | | | ele | | | | | 1 |

Um primeiro aspecto a notar na Tabela 2 é a prevalência de espaços *vazios*. Note como a maioria absoluta das células ficaram em branco: chamamos isso de *esparsidade*. Veja como há pouca sobreposição entre as palavras-alvo, no que diz respeito aos itens de contexto com os quais co-ocorrem. Só observamos alguma sobreposição nas primeiras linhas da tabela que correspondem – e isso não é coincidência – aos itens mais frequentes do corpus como um todo. Essa esparsidade é parte da natureza do problema: as línguas naturais obedecem à Lei de Zipf, segundo a qual, ao ordenar por frequência e em ordem decrescente as palavras de uma amostra significativa de textos (ou de fala), a frequência de um dado item na posição *p* é aproximadamente igual à frequência do primeiro item da lista dividida por *p*. Assim, uma pequena quantidade de itens é muito frequente e a grande maioria ocorre pouquíssimas vezes. Em decorrência disso, apenas um número restrito de itens tem frequência suficiente para co-ocorrerem com a maioria das palavras da amostra, sendo assim mais úteis para sua classificação distribucional na medida em que maximizam o preenchimento da matriz de co-ocorrência.

**4. A construção de um modelo computacional**

A metodologia adotada no exemplo que segue é a da modelagem computacional, como explicada anteriormente. Em nossos estudos, utilizamos corpora de fala espontânea dirigida à criança e também dados de fala entre adultos. Ter esses dois conjuntos de dados à disposição nos permite, por exemplo, avaliar diferenças possíveis entre essa fala dirigida à criança, descrita como “manhês” (ou “paiês”), e a fala típica entre falantes adultos. Estudos como o de Redington et al.

(1998) para o inglês e os de Faria e Ohashi (2018) e Faria (2019a, 2019b), para o português brasileiro (PB), partem do tipo de análise distribucional proposta de Harris (op.cit.) e ilustrada na seção anterior, para investigar o grau de *informatividade* da informação distribucional.[7] O que isso quer dizer? O intuito é mostrar se tais informações são úteis na aprendizagem de categorias sintáticas. Sendo úteis, isso justificaria que as várias abordagens teóricas levassem em conta o papel da informação distribucional no processo de aquisição da linguagem e a incluíssem em seus modelos teóricos. Esse é um exemplo do tipo de impacto que se espera de modelagens computacionais.

Ao abordar um problema de aquisição computacionalmente, é bastante útil se guiar por questões orientadoras da especificação, construção e avaliação dos modelos, tais quais as apresentadas por Bertolo (2001), que visam estabelecer critérios objetivos para "traduzir" um problema de aquisição para um problema de modelagem computacional. Cada questão incide sobre aspectos particulares do problema, a saber: as propriedades do conhecimento adquirido, os procedimentos de aquisição, as propriedades dos dados de entrada e de sua apresentação, as restrições sobre os mecanismos de aprendizagem e a condição de sucesso ou convergência (p.e., percentual mínimo de itens lexicais adquiridos). Vejamos então como respondemos, para modelar a aprendizagem distribucional, as questões de Bertolo (op.cit.):

1) O que é aprendido, exatamente?
    - Pode-se dizer que conhecer as categorias sintáticas das palavras da língua é ser capaz de *determinar, para um par qualquer de palavras conhecidas, se elas pertencem ou não à mesma categoria*. Esse é, portanto, o conhecimento a ser adquirido. Note que precisamos, portanto, estabelecer de antemão quais seriam as categorias a ser aprendidas. Essa decisão é absolutamente determinada pela teoria (ou teorias) que assumimos, sendo assim uma variável do problema.
2) Quais tipos de hipóteses o aprendiz computacional é capaz de entreter?
    - A estrutura base das hipóteses é: *cada palavra tem apenas uma categoria possível e palavras de mesma categoria têm distribuições de contexto similares*. Esta é sem dúvida uma definição simplificadora: algo pervasivo nas línguas naturais é o fato de que uma mesma forma lexical pode corresponder a duas ou mais categorias (homofonia). Na definição acima, hipóteses podem variar quanto ao contexto

[7] A tradição de investigações sobre aprendizagem distribucional já é longa e datamos seu início com Finch e Chater (1991), embora com trabalhos precursores (ver Turney e Pantel, 2010).

assumido. O grau de similaridade requerido também é uma variável, mas nesse estudo ela é determinada automaticamente.

3) Como os dados da língua-alvo são apresentados ao aprendiz?
    - O aprendiz recebe enunciados (de fala espontânea) transcritos, já segmentados em palavras e sem pontuação intra-sentencial. Os dados são apresentados *todos de uma vez* para processamento.
4) Quais restrições governam o modo como o aprendiz atualiza suas conjecturas em resposta aos dados?
    - Como palavras são avaliadas com base em seus contextos, o aprendiz atualiza suas conjecturas agregando informação estatística sobre as palavras, na forma de *vetores de contexto*, isto é, vetores que contém as *frequências de co-ocorrência entre palavras-alvo e palavras de contexto*. Parâmetros que podem variar aqui são a janela de contexto, a quantidade de palavras mais frequentes que são itens contextuais válidos e o uso (ou não) de fronteiras entre enunciados.
5) Sob quais condições, exatamente, dizemos que o aprendiz obteve sucesso na tarefa de aprendizagem da linguagem?
    - Quando a *classificação postulada* pelo aprendiz é significativamente mais parecida com a *classificação de referência* do que a *classificação aleatória* das palavras.

As respostas acima tornam explícitos os componentes básicos do modelo e também evidenciam assunções, simplificações (note que fizemos várias, em relação ao fenômeno empírico) e escolhas globais de modelagem. No entanto, ainda não há aí um modelo propriamente dito e nem um procedimento de aprendizagem. Na verdade, cada uma das respostas produz consequências para o modelo a ser construído, como vemos adiante. A tarefa seguinte é, portanto, especificar o modelo computacional que, nesse caso, se caracteriza por um algoritmo que implementa o método de aprendizagem distribucional. Esse método consiste de três estágios para aprendizagem distribucional das categorias a partir de dados de fala dirigida à criança, que podem ser obtidos, por exemplo, na base CHILDES (MACWHINNEY, 2000):

(i) Medir os contextos de distribuição em que cada palavra ocorre;
(ii) Comparar o contexto de distribuição para pares de palavras;
(iii) Agrupar palavras com distribuições similares.

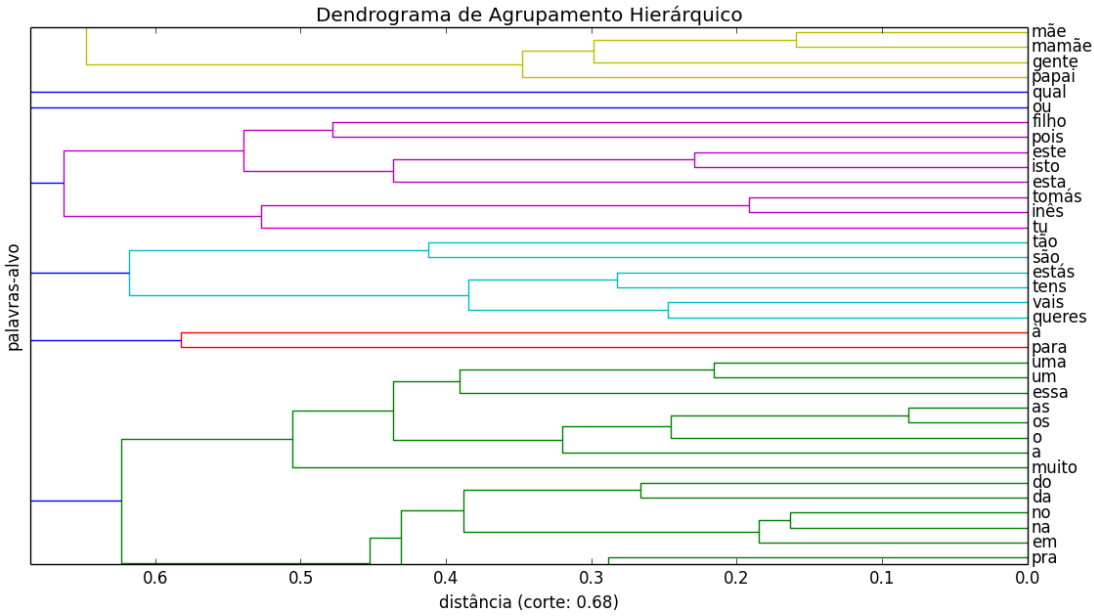
A etapa (i) envolve medir o contexto em que as palavras ocorrem, atendendo mais diretamente aos requisitos das respostas (2) e (4) acima. Além disso, a decisão em (3) afeta diretamente as propriedades e o modo como esta etapa vai operar. Como o aprendiz processa o corpus todo de uma vez, podemos extrair dos dados o ranqueamento de palavras por frequência. Daí, tiram-se as 1000 palavras mais frequentes para serem categorizadas e as 150 mais frequentes para serem itens contextuais válidos. Esses recortes são importantes para limitar o problema de esparsidade indicado anteriormente: itens mais frequentes aumentam bastante a chance de preenchermos a maior parte da tabela (o que aumenta o poder discriminatório do modelo). Assim, são coletadas estatísticas de co-ocorrência entre palavras-alvo e as palavras (válidas) em seu entorno, armazenando essa informação em uma matriz de co-ocorrência (ver Tabela 2, já apresentada). Nesta, cada linha representa a distribuição de uma dada palavra-alvo e as colunas representam as frequências das palavras contextuais em relação a ela. Assim, a tabela pode ser vista como uma lista de *vetores de contexto* das palavras.

Em seguida, na etapa (ii), avalia-se a similaridade entre esses vetores de contexto, que podem ser pensados como pontos em um espaço multidimensional de possíveis distribuições de palavras. Essa etapa vai fornecer as bases sobre as quais é possível atender às respostas (2) e (5): sem o cálculo da similaridade para todas as combinações em pares das palavras-alvo, não é possível agrupá-las e, por conseguinte, não se obtém classificações. Conforme (2), espera-se que palavras de mesma categoria sintática tenham distribuições similares, ou seja, estejam relativamente próximas nesse espaço multidimensional. Para calcular a similaridade, no exemplo que trazemos aqui, utiliza-se o *coeficiente de correlação de postos de Spearman*, que é uma dentre várias métricas de similaridade alternativas (p.e., cosseno entre vetores e a distância euclidiana). Aqui vale observar que a decisão quanto à métrica de similaridade é um exemplo de aspecto arbitrário da modelagem, visto que, nesse caso, não decorre de nenhuma teoria psicolinguística sobre a natureza do processamento estatístico da linguagem ou, mesmo, da cognição de modo geral. É mais comum do que gostaríamos, nas modelagens, nos depararmos com situações em que simplesmente não há uma diretriz teórico-empírica sobre um dado aspecto do modelo. Nesses casos, o melhor que podemos fazer é explorar alternativas (quando houver) para determinar quais delas produzem os melhores resultados, o que, em nosso caso, implica maximizar o aprendizado.

Finalmente, na etapa (iii), os agrupamentos são gerados usando o método de *análise de agrupamento hierárquico padrão* (Redington *et al.*, 1998, p. 437), outra decisão arbitrária, visto

que há alternativas matemáticas para a obtenção dos agrupamentos e não há, por outro lado, estudos que avaliem sua plausibilidade psicológica. Nessa etapa, grosso modo, o algoritmo começa agrupando as duas palavras mais próximas (por similaridade) e segue adiante, agrupando outras palavras e grupos entre si, encerrando quando um único grupo final é obtido. O agrupamento hierárquico gerado desta forma pode ser representado como um dendrograma, como mostra a Figura 1.

Figura 1 – Dendrograma parcial de agrupamentos de palavras.




O resultado dessa etapa é produzir a base sobre a qual o aprendiz computacional pode extrair os agrupamentos (classificação postulada) que mais se aproximem da classificação de referência, que foi anotada e revista por humanos. Essa extração, no modelo, envolve traçar um corte vertical no "ponto ótimo" do dendrograma acima, de modo que os agrupamentos obtidos na altura do corte são aqueles que mais se aproximam da referência. Com isso, atendemos às questões (1) e (5): ao ser capaz de determinar se duas palavras são (ou não) da mesma categoria, pode-se avaliar o (grau de) sucesso do aprendiz na aprendizagem.

Um último aspecto importante na construção de modelo é dar um suporte adequado a qualquer hipótese sobre quais propriedades devem ser incluídas nos dados de entrada e quais devem ser deixadas de fora. Nesse sentido, Morgan (1986) argumenta que, para estabelecer a plausibilidade de qualquer hipótese sobre as informações presentes nos dados de entrada, é necessário que as seguintes condições empíricas sejam satisfeitas:

i. *Universalidade*: a informação deve estar universalmente presente.
ii. *Disponibilidade*: a informação deve ser acessível ao aprendiz.
iii. *Representabilidade*: o aprendiz deve ser capaz de representá-la apropriadamente.

iv. *Papel identificável*: a informação deve ter um papel identificável na aquisição.

v. *Necessidade*: crucialmente, mesmo atendendo aos quatro critérios acima, é preciso demonstrar que essa informação é realmente necessária para a aquisição e não apenas facilitadora.

Como vimos, a resposta (3) indica que os dados de entrada se compõem basicamente de enunciados transcritos e pré-segmentados em palavras. Essa, como se vê, é uma espécie de assunção mínima que deixa de fora informações morfológicas detalhadas, informações fonológicas (incluindo a prosódia), informações semânticas e também sintáticas (p.e., fronteiras de sintagmas). Ou seja, é provável que o acréscimo de qualquer dessas informações leve a uma significativa melhora na performance do aprendiz computacional. Nesse sentido, essa representação atenderia a todos os critérios acima (com base em JUSCZYK, 2000, e HÖHLE, 2009): (i) os enunciados provenientes do corpus de aquisição representam dados universalmente disponíveis (em condições típicas), (ii) ao assumir dados já segmentados, assume-se que o aprendiz modela uma criança do segundo ano de vida em diante, (iii) sabemos que a criança é capaz de atribuir algum tipo de representação aos itens lexicais (pois os reconhece já ao final do primeiro ano de vida), (iv) crianças são sensíveis à ordem das palavras nos enunciados, e (v) sem enunciados transcritos que mantém a informação da ordem de palavras da língua, a informação distribucional não pode ser extraída. Por outro lado, a forma de apresentação dos dados é implausível: não é o caso de que crianças processam toda sua experiência linguística de uma vez. Ao contrário, o que se observa é uma aprendizagem gradual, incremental, em que a criança processa e aprende a partir de cada enunciado, um por vez.

## 5. Possibilidades exploratórias que um modelo abre

Um modelo como o apresentado acima permite que se investigue uma série de aspectos do problema, a partir da manipulação e do controle de suas variáveis. Dos exemplos que daremos abaixo, praticamente nenhum seria viável de estudar experimentalmente com crianças. Isso confere à modelagem computacional um estatuto de ferramenta complementar fundamental para uma área que pretenda de fato aprofundar seu conhecimento sobre a aquisição da linguagem. Como exemplo, tomando os estudos em Redington et al. (1998) e sua reprodução para o português em Faria e Ohashi (2018) e Faria (2019a, 2019b), temos um número expressivo de

experimentos possíveis, apresentados a seguir com uma breve indicação dos resultados alcançados:

*Variação de janelas de contexto*

Com o modelo, pode-se manipular a janela de contexto para avaliar quais posições contextuais são (mais) informativas para aprendizagem distribucional na língua em questão. Assim é possível comparar como línguas distintas se comportam a esse respeito. Por exemplo, uma avaliação possível seria comparar a informação das duas palavras anteriores à informação das duas palavras posteriores, com respeito a uma dada palavra-alvo. Isso permitiu aos estudos citados atestar, por exemplo, que a janela (relativa) [-2, -1, 1] (i.e., duas palavras anteriores e uma palavra posterior) é a mais informativa para o PB, enquanto para o inglês[8] seria [-2, -1, 1, 2]. Além disso, em ambas as línguas, palavras precedentes são mais informativas quanto à categoria da palavra.

*Variação do número de palavras-alvo e de contexto*

É possível também variar o número de palavras-alvo e de contexto utilizadas. Com isso, pôde-se observar, por exemplo, que analisar muitas palavras-alvo pouco frequentes leva a uma queda na *performance* geral do aprendiz. Ao mesmo tempo, aumentar o número de palavras de contexto utilizadas só ajuda até certo ponto: a partir de 100 palavras de contexto, em PB, a *performance* cai gradativamente, o que pode indicar que a criança não rastreia toda e qualquer relação distribucional, mas talvez apenas aquelas mais recorrentes. Ademais, em um modelo que assume apresentação "instantânea" aos dados, a possibilidade de manipular esses parâmetros oferece um possível caminho aproximativo para estudar como a gradualidade da exposição afetaria o comportamento do aprendiz.

*Avaliação do valor da informação distribucional por categoria*

Uma vez que o vocabulário inicial da criança mostra preponderância de substantivos, é possível que certas categorias sejam mais fáceis de aprender (distribucionalmente) do que outras. Com o modelo podemos observar a performance do aprendiz em cada uma das

[8] Nossos próprios resultados para o inglês, no entanto, revelam que a janela [-1, 1] é que produz a melhor performance, diferentemente dos resultados de Redington et al. (1998). Os dados utilizados foram, até onde podemos saber, praticamente os mesmos do estudo original.

categorias sintáticas assumidas. Os resultados indicam que categorias "abertas" ou "de conteúdo" (verbos, substantivos, adjetivos e advérbios) estão mais ao alcance do método distribucional do que as funcionais.

*Variação da quantidade de dados de entrada*

Uma informação importante para um modelo como este é qual a quantidade mínima de dados necessários para que o método seja efetivo. Neste experimento, são utilizadas porções crescentes do corpus, até chegar ao corpus inteiro. O que observamos para o PB é que o método não depende de muitos dados (relativamente ao que se estima que uma criança esteja exposta) para se mostrar efetivo.

*Uso de fronteiras de enunciado*

Será que estar na fronteira do enunciado é relevante para classificar uma palavra? Neste experimento, fronteiras são marcadas explicitamente no corpus e usadas na classificação. Os resultados mostram uma melhora na performance do aprendiz, neste caso. Um próximo passo natural seria estudar fronteiras prosódicas, exatamente o que faz Mintz et al. (2002) de forma aproximada. Um corpus anotado com fronteiras de frase fonológica, por exemplo, permitiria um estudo ainda mais preciso dessa questão.

*Comparação de frequências vs. ocorrência*

Como comentado na seção 3, apenas atestar a co-*ocorrência* de uma dada palavra-alvo com uma palavra de contexto não parece ser suficiente para explorar ao máximo a informação distribucional. Com este experimento, isso é averiguado plenamente e os resultados indicam que a frequência é de fato importante para um melhor aprendizado.

*Remoção de palavras funcionais*

Itens funcionais auxiliam ou atrapalham a classificação das palavras de conteúdo? Dado que tais itens são mais tardios na fala da criança, poderia ser o caso de que elas os ignorassem inicialmente para fins de classificação de palavras. Para avaliar isso, este experimento remove as palavras funcionais do corpus e avalia a performance do aprendiz. Os resultados mostram, no entanto, que estes itens são fundamentais para uma boa performance do aprendiz. Tal resultado é compatível com outros estudos que mostram

que a criança utiliza perceptivamente essa informação, mesmo quando sua produção não reflete isso diretamente (ver, p.e., CHRISTOPHE, MILLOTTE, BERNAL e LIDZ, 2008).

*O efeito de uma categoria sobre a aquisição de outras*

Nas simulações anteriores, apenas palavras específicas foram consideradas como itens contextuais. Porém, será que o aprendiz se beneficiaria em utilizar a categoria (já adquirida) de uma palavra como informação contextual, ao invés da própria palavra? O modelo permite que itens sejam substituídos pela sua própria categoria (pré-definida) no corpus (p.e., itens funcionais todos substituídos por FUNC) e o modelo é então reavaliado. Este tipo de estudo permite avaliar detalhadamente o efeito de cada categoria particular sobre as demais e até mesmo de mais de uma categoria ao mesmo tempo, se assim quisermos. Os resultados mostram que não há efetiva melhora, mas que a categoria das palavras funcionais é mais informativa (para aquisição das categorias de substantivos, verbos, adjetivos e advérbios) do que as categorias dos substantivos ou dos verbos para as demais.

*A fala dirigida à criança facilita a aprendizagem?*

Finalmente, uma vez que se tenha à disposição um corpus com dados de fala entre adultos, pode-se fazer a comparação das performances do modelo quando usa um ou outro tipo de fala. A possibilidade de fazer um experimento em larga escala desse tipo complementa análises mais qualitativas das diferenças entre o "manhês" e a fala típica. Os resultados para ambas as línguas indicam que não haveria benefícios em termos de informação distribucional na fala dirigida à criança.

Até aqui, buscamos ilustrar a gama de possibilidades que um modelo computacional abre para avaliar em larga escala aspectos da aquisição, o que seria impraticável com crianças em estudos experimentais. Ademais, a partir desses exemplos, é possível pensar em várias outras manipulações, tais como inserir gradativamente informação morfológica nas análises, avaliar diferentes classificações de palavras (p.e., de caráter mais formal *vs*. funcional ou comunicativo, assunção de grandes categorias *vs*. categorias mais subdivididas), avaliar o efeito de fronteiras de sintagma possivelmente coincidentes com fronteiras de frase fonológica (usando, para isso, um corpus anotado sintaticamente), entre outras coisas. Finalmente, quanto mais aspectos do

problema o modelo incluir em sua arquitetura, mais possibilidades de estudos se abrem.

**6. Considerações finais**

Como Schunk (2012, p.11) comenta, muitas vezes não temos ainda uma compreensão clara sobre como as várias variáveis envolvidas em um fenômeno o afetam e interagem entre si. Um modelo computacional permite, no entanto, configurar experimentos em que avaliamos cada variável isoladamente e também em sua interação com quaisquer outras. Ou seja, permite tanto estudos *correlacionais*, para investigar interações entre variáveis, quanto estudos *experimentais*, para investigar possíveis relações de causa e efeito entre elas. O potencial exploratório que isso abre para as investigações em aquisição é significativo e de valor inestimável para uma área que precisa integrar as várias metodologias de investigação para ampliar suas possibilidades de compreensão da complexidade e da dinâmica que caracterizam o processo de aquisição da linguagem.

Neste artigo, buscou-se desenvolver no leitor leigo em modelagens computacionais um olhar de "modelador", deixando de lado aspectos mais técnicos e focando no processo que vai desde a descrição de um problema de aquisição, passando pela sua "tradução" para um problema de modelagem e chegando ao desenvolvimento de um modelo propriamente dito. Neste percurso, tomamos consciência de como a modelagem impõe escolhas concretas, precisas e explícitas por parte do pesquisador, de modo que suas assunções, simplificações e decisões arbitrárias fiquem evidentes, sejam elas em parte determinadas pela própria teoria ou não. Nesse sentido, não é demasiado destacar novamente a importante distinção entre o/um modelo e a realidade empírica: sempre há simplificações em todos os aspectos envolvidos. Quando falamos em aprendiz computacional nos referimos a um simulacro do que seria a criança em relação ao processo em questão. Do mesmo modo, quando falamos em aprender categorias sintáticas e em dados de entrada, estamos também nos referindo a aproximações que, apesar do termo, podem estar ainda muito distantes da realidade. É preciso termos sempre uma consciência plena disso, para que nosso olhar sobre o modelo seja arguto e equilibrado, sem um otimismo ingênuo, que tomaria o modelo como reflexo definitivo da realidade, mas também sem o ceticismo obtuso que negaria em princípio qualquer contribuição possível às modelagens.

Assim, esperamos ter contribuído aqui, mesmo que de modo incipiente, para evitar tais extremos. Uma vez que todos compreendamos esse processo que leva aos modelos

computacionais, nós, enquanto comunidade que investiga a aquisição, nos tornaremos mais capazes de avaliar criticamente os modelos e, mais que simplesmente apontar limitações, apontar também caminhos para seu aprimoramento ou para a plena exploração de seu potencial, através da elaboração de novos experimentos que respondam a novas perguntas.

**Referências**

BERALDO, R. L. (2020). *Computational models of lexical acquisition: surveying the state of the art* (Dissertação de Mestrado). Universidade Estadual de Campinas, Campinas, Brasil.

BERNAL, S., LIDZ, J., MILLOTTE, S., e CHRISTOPHE, A. (2007). Syntax constrains the acquisition of verb meaning. *Language Learning and Development*, 3, 325–341.

BERTOLO, S. (2001). A brief overview of learnability. In ___ (ed.). *Language acquisition and learnability*. Cambridge: Cambridge University Press, p. 1–14.

BORGES NETO, J. O pluralismo teórico na linguística. In: __. *Ensaios de filosofia da linguística*. São Paulo: Parábola Editorial, 2004, p. 67-93.

CHRISTOPHE, A., MILLOTTE, S., BERNAL, S., e LIDZ, J. (2008). Bootstrapping Lexical and Syntactic Acquisition. *Language and Speech*, 51 (1&2), 61–75.

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

FRANK, M. C. (2011). *Computational models of early language acquisition*. Online. URL: http://langcog.stanford.edu/papers/F-underreview-b.pdf. Acesso em 15/05/2020.

GOODLUCK, H. (1991). *Language Acquisition: A Linguistic Introduction*. Blackwell, Oxford.

GUASTI, M. T. (2002). *Language acquisition: a linguistic perspective*. Cambridge, Mass: The MIT Press.

HÖHLE, B. (2009). Bootstrapping mechanisms in first language acquisition. *Linguistics*, 47-2, 359–382.

JUSCZYK, P. W. (2000). How Speech Perception Develops During the First Year. In __, *The Discovery of Spoken Language*. Cambridge, MA: The MIT Press, pp. 73–109.

KAPLAN, F., OUDEYER, P.-Y., & BERGEN, B. (2008). Computational models in the debate over language learnabiliy. *Infant and Child Development*, 17(1):55–80.

LANDAU, B., e GLEITMAN, L. R. 1985. *Language and experience: Evidence from a blind child.* Cambridge, Mass.: Harvard University Press.

MACWHINNEY, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ, third edition edition.

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.

MORGAN, J. L. (1986). *From simple input to complex grammar*. The MIT Press.

NAIGLES, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language* 17, 357-374.

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). 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. (2019b). 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

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

PINKER, S. (1979). Formal models of language learning. *Cognition*, 7:217–283.

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.

SCHUNK, D. H. (2012). Introduction to the Study of Learning. In ___, *Learning theories: an educational perspective*. Boston, MA: Pearson Education, Inc., pp. 1–28.

SEIDENBERG, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, 275(14):1599–1603.

TOMASELLO, M. (2009). The usage-based theory of language acquisition. In Edith L. Bavin (Ed.), *The Cambridge handbook of child language*. Cambridge: Cambridge Univ. Press, pp. 69–87.

TURNEY, P. D. e PANTEL, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

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

YANG, C. (2011). Computational models of syntactic acquisition. *WIREs Cogn Sci,* doi: 10.1002/wcs.1154