UM ESTUDO SOBRE APRENDIZAGEM DISTRIBUCIONAL DE CATEGORIAS SINTÁTICAS NO PORTUGUÊS BRASILEIRO

Giulia Osaka OHASHI
Orientador: Prof. Dr. Pablo Faria

RESUMO: Em Redington et al. (1998), o potencial da informação distribucional na categorização lexical do inglês é analisado com base em uma série de experimentos computacionais. Tomando-o como base, propomos replicar experimentos e análises conduzidos ali sobre dados do português brasileiro (PB), contribuindo para uma avaliação translinguística. Os corpora consistirão em dois conjuntos de dados: dados de fala dirigida à criança e dados de diálogos entre adultos. Os primeiros serão compilados a partir da Coleção "Projeto de Aquisição da Linguagem Oral" e dos dados do PB disponíveis na base CHILDES. O segundo conjunto será obtido na plataforma NURC ("Projeto Norma Linguística Urbana Culta - RJ").
Palavras-chave: aquisição da linguagem; aprendizagem distribucional; modelo computacional; categoria sintática; português brasileiro.

1. INTRODUÇÃO

Como se dá o processo de aquisição das línguas humanas por crianças? O que explica uma aquisição rápida, uniforme e espontânea, como a que se verifica empiricamente? Quais estratégias de aprendizagem são utilizadas pela criança? São perguntas como estas que movem a área de aquisição da linguagem.

Neste trabalho, é apresentada uma pesquisa que visa contribuir nesse campo ao se debruçar sobre o processo de categorização de palavras, à medida em que a criança adquire vocabulário. Mais especificamente, tendo por base o estudo apresentado em Redington et al. (1998), propomos investigar computacionalmente o potencial da informação distribucional como fonte de informação para a classificação de palavras em classes lexicais no português brasileiro (PB), sobre dados retirados das bases CHILDES (MACWHINNEY, 1989), da Coleção "Projeto Aquisição da Linguagem Oral[1]" e do Projeto NURC/RJ[2] ("Projeto Norma Linguística Urbana Culta - RJ"). Ao analisar dados do PB, este trabalho visa contribuir para uma avaliação translinguística dos achados apresentados em Redington et al. de modo a, conforme os resultados encontrados, dar

1. Inventário do Projeto Aquisição da linguagem Oral. Org. Vania Regina Personeni. Campinas: CEDAE/IEL, s.d. 33p.; Disponível em: <https://goo.gl/Jop0NU>. Acesso em: 27 maio 2017.

2. Disponível em: <http://www.letras.ufrj.br/nurc-rj/>. Acesso em: 27 abr. 2017.

suporte às conclusões daquele estudo ou trazer à tona contra-evidências que ensejem novas investigações sobre esse aspecto da aquisição da linguagem.

2. A TEORIA DE AQUISIÇÃO DA LINGUAGEM

A aquisição da linguagem é um processo natural do desenvolvimento humano típico, ou seja, todas as crianças típicas expostas a um ambiente onde exista comunicação linguística adquirirão a língua em um mesmo período de tempo e fluentemente. Para estudar e, assim, entender melhor esse processo, o campo da aquisição da linguagem dispõe de diferentes metodologias e abrange várias orientações teóricas. Neste trabalho, assumimos o viés gerativista, em que a aquisição da linguagem é vista de uma perspectiva inatista e formal.

2.1. Gramática gerativa

Inaugurada por Noam Chomsky nos anos 1950, a Gramática Gerativa tem se mostrado bastante frutífera nos estudos linguísticos. Nela há a proposta de uma Gramática Universal (GU) (COSTA & SANTOS, 2003), que seria inata à espécie humana e incorporaria princípios e propriedades comuns a todas as línguas naturais. Segundo Chomsky (1965 *apud* INGRAM, 1989), o papel do linguista é descrever a *competência* linguística, isto é, o sistema de regras subjacente que todo falante nativo de uma língua possuiria em sua mente. Chomsky o diferencia da *performance*, que diz respeito ao "uso" desse conhecimento na produção e na compreensão de enunciados da língua. Algumas evidências reforçam a visão da linguagem como dotação biológica do ser humano, como, por exemplo, o fato de as línguas naturais serem específicas à espécie humana e/ou elas exibirem propriedades características do desenvolvimento biológico, tais como o desenvolvimento da língua ser espontâneo, exitoso em todos os indivíduos típicos da espécie e uniforme entre as crianças de uma comunidade. Outrossim, segundo Costa & Santos (2003), certos comportamentos infantis reforçam a teoria inatista: as crianças dizem o que nunca ouviram (inovam), são sistemáticas nos seus erros e estes normalmente revelam um certo conhecimento da gramática. Tais características, entre outras, conferem à abordagem inatista uma significativa relevância nos estudos de aquisição.

Outro importante argumento chomskyano para a hipótese da dotação inata é o da 'pobreza de estímulos' (GROLLA & FIGUEIREDO SILVA, 2014). Segundo este, ainda não há uma resposta alternativa científica satisfatória para o problema de como a criança adquire o conhecimento gramatical. As crianças o fazem em um curto período de tempo, com dados incompletos e degradados (contendo reformulações, interrupções, barulhos etc.), input aleatório (não há uma sistematização do ensino da língua para crianças) e sem

evidências negativas (isto é, quando uma criança produz algo errado, normalmente não há correção, e, caso haja, a criança em geral ignora as correções). Além disso, deve-se notar que a criança está em fase de desenvolvimento e não está "preocupada" somente em aprender a falar, mas sim a comer, andar etc. Como a criança não presta atenção a todas as informações de modo uniforme, entra a questão de quais seriam os tipos de informação que a criança privilegia (COSTA & SANTOS, 2003). Como já dito acima, sendo a aquisição uniforme e espontânea, muitos estudiosos defendem que a aquisição da linguagem seria praticamente impossível sem uma dotação inata específica, como defende a abordagem racionalista[3].

Como exemplo da incompletude dos dados a que a criança está exposta, Grolla & Figueiredo Silva (2014) apresentam o exemplo do 'você' que pode ser encurtado como 'cê' em alguns contextos sintáticos. Embora 'você' possa ser usado tanto como sujeito ("você contou para ele"), quanto como objeto da oração ("ele contou para você"), não acontece o mesmo com 'cê': "cê contou para ele[4]" é gramatical, mas não "ele contou para cê" . Note que não nos é ensinado explicitamente sobre essa regra, isto é, a criança só tem acesso a dados positivos (sentenças gramaticais), mas mesmo assim chega a saber quais são as impossibilidades gramaticais da língua. Esse problema é também chamado na literatura de "problema da projeção", isto é, como "projetar" uma gramática capaz de produzir e interpretar infinitas expressões a partir de um conjunto finito de exemplos:

> "Assim, se é verdade que os dados linguísticos primários[5] são necessários, parece claro também que eles não são suficientes para chegarmos a tudo o que caracterizamos como o conhecimento de uma língua, e, portanto, algum tipo de mecanismo de outra ordem é necessário para responder por esse 'pulo do gato' que a criança dá." (GROLLA & FIGUEIREDO SILVA, 2014, p.44).

Com o intuito de desvendar essas questões, o gerativismo enfatiza dois pontos centrais para a pesquisa em aquisição da linguagem: (i) o tipo de dado usado para se fazer afirmações sobre a linguagem infantil e (ii) a diferença entre a competência e a performance da criança (INGRAM, 1989). Com relação ao primeiro, Brown (1973 *apud* INGRAM, 1989) e Miller & Ervin (1964 *apud* INGRAM, 1989) propõem complementar o dado de fala espontânea da criança (utilizado desde o início dos estudos sobre a aquisição

3. A visão racionalista postula que a criança nasce com conhecimentos especificamente linguísticos. A exposição a uma determinada língua, portanto, faria com que a criança moldasse seus conhecimentos de acordo com sua língua-alvo (Grolla & Figueiredo Silva, 2014).

4. Perceba que seria possível ('ele contou procê'), mas, nesse caso, temos somente a elisão do 'a' e do 'v' nas palavras 'pra' e 'você', respectivamente; sendo portanto, ainda a utilização do 'você' como objeto.

5. "'Dados linguísticos primários' são as sentenças que os falantes produzem ao redor da criança, dirigindo-se ou não a ela. Ou seja, é tudo aquilo que a criança ouve ao seu redor". (GROLLA & FIGUEIREDO SILVA, 2014, p.44). Tais dados, ou 'input', fornecem informações sobre a língua materna que a criança está adquirindo.

da linguagem) com outros tipos de procedimentos para obter amostras da língua, tais como testar a compreensão ou elicitar a imitação.

2.2 Corpora de aquisição

Em qualquer projeto de pesquisa, o primeiro passo é a revisão mais aprofundada da literatura (INGRAM, op.cit.). Em aquisição, temos um passo adicional, a avaliação cuidadosa de dados relevantes nos acervos disponíveis. Ingram (1989) afirma que o campo da aquisição da linguagem da criança precisa ter as seguintes propriedades: (i) requer o estudo da linguagem infantil; (ii) examinar os dados de crianças com teorias gramaticais bem definidas; e (iii) concentrar nos métodos que estabelecem quando o comportamento linguístico da criança é baseado em regras. Tendo em vista esse escopo, para Grolla & Figueiredo Silva (2014), a coleta de dados para a investigação no campo da aquisição da linguagem pode ser dividida em dois tipos: a coleta de dados espontâneos, isto é, não guiar a criança a falar de algo ou algum modo específico, e a coleta de dados experimentais, que, ao contrário do espontâneo, controla as condições de fala para dar julgamentos sobre sentenças apresentadas a ela.

Os dados espontâneos são obtidos através de gravações da fala de criança, podendo ter por volta de 30 minutos cada sessão, em um período que pode variar mais ou menos de um até cinco anos. Nas gravações feitas neste método, utiliza-se uma transcrição uniformizada, para que a pesquisa em cima desses dados seja facilitada. A produção espontânea oferece um panorama amplo sobre as estruturas que a criança gravada utiliza, no entanto, esse método é limitado, visto que alguns detalhes podem não aparecer, como Grolla & Figueiredo (op.cit.) exemplificam com o caso da voz passiva (estrutura pouco utilizada até pelos adultos). Ou seja, a criança pode entender e produzir algumas construções linguísticas e essas não aparecerem nas gravações espontâneas.

Segundo Grolla & Figueiredo (op.cit.), os dados experimentais então complementam as informações obtidas para a pesquisa, pois o investigador pode testar essas construções em experimentos desenhados especialmente para esse objetivo. Com crianças com mais de três anos de idade, na produção é possível elicitar estruturas em contextos lúdicos fazendo perguntas para crianças ou podemos pedir para que a criança imite certas construções, o que pode ser indicativo de o conhecimento daquelas estruturas. Já na compreensão, são utilizadas quatro tipo de tarefas: (a) Tarefa de Julgamento de Valor de Verdade; (b) Tarefa de Julgamento de Gramaticalidade; (c) Tarefa de encenação; e (d) Tarefa de escolha de figuras. Vale ressaltar que a escolha de qualquer um desses métodos depende muito do tipo de fenômeno gramatical que está sob observação.

Por conseguinte, segundo ainda os mesmos autores, o método experimental combinado com o espontâneo fornece uma ideia mais completa e precisa sobre o

desenvolvimento infantil. Ademais, uma vantagem prática do método experimental é a possibilidade de obter dados de um número maior de crianças do que pelo método espontâneo. Vê-se dessa forma que, cabendo à teoria investigar as propriedades e mecanismos envolvidos no processo de aquisição da linguagem, são necessários diferentes métodos de averiguação das teorias linguísticas para os mais diversos tipos de fenômenos gramaticais. No estudo em andamento apresentado aqui, em função de seus objetivos, utilizamos dados espontâneos de fala dirigida à criança e de diálogos entre adultos, como detalhamos mais adiante.

3. MODELOS COMPUTACIONAIS COMO MEIO DE INVESTIGAÇÃO

Um outro meio de investigação de questões colocadas pelos estudos de aquisição é o desenvolvimento de modelos computacionais que incidam sobre essas questões. Tais modelos computacionais são tentativas de aproximação dos processos psicolinguísticos e ambientais que se dão em crianças desde os primeiros estágios da aquisição da linguagem (PEARL, 2010). Na modelagem computacional (MARR, 1982 *apud* PEARL, 2010), é necessário lidar com três níveis de processamento de informação, sendo os dois primeiros responsáveis por tratar de questões de modelagem, enquanto o terceiro trata da engenharia da construção do modelo.

Primeiramente, o nível computacional, relativo à descrição do problema, é o que mais dialoga com a teoria psicolinguística, pois é nesse, através de modelos teóricos, que os mais diversos aspectos da linguagem devem ser especificados. Secundariamente, o nível algorítmico, relativo aos passos necessários para a solução dos problemas de aprendizagem, especifica os procedimentos de aquisição responsáveis por induzir ou deduzir a gramática a partir dos dados de entrada. Neste nível, define-se aspectos do paradigma de aprendibilidade tais como, por exemplo, se a indução de gramáticas será ou não baseada em evidência negativa direta ou indireta (FARIA, 2013). E, finalmente, temos o nível implementacional em que o modelo é implementado computacionalmente. Neste nível, verifica-se a capacidade de tais modelos de refletirem observações empíricas do processo de aquisição e de fazerem novas previsões sobre o processo (PEARL, 2010).

Uma das principais virtudes dos modelos computacionais é a necessidade de explicitude quanto às suas assunções, propriedades e mecanismos propostos ou assumidos; algo nem sempre encontrado na teoria linguística. Segundo Kaplan (2007), os modelos servem tanto para investigar problemas que não estão resolvidos, como para validar hipóteses teóricas já propostas, buscando se aproximar cada vez mais em suas simulações dos resultados da vida real. Por esta razão, para que simulações sejam consideradas plausíveis e substanciais, são necessárias algumas condições pertinentes às modelagens. O embasamento teórico, por exemplo, é crucial para investigar uma afirmação específica

sobre aquisição da linguagem (PEARL, 2010). Consequentemente, esta virtude tem seu efeito colateral: os modelos computacionais serão no mínimo tão limitados quanto as teorias em que se baseiam.

Finalmente, modelos devem se avaliados segundo certos critérios que, na visão de Pinker (1979), seriam:

i. Aprendibilidade: o aprendiz deve aprender o que é esperado;
ii. Equipotencialidade: deve aprender em qualquer língua;
iii. Entrada: deve aprender com os dados de entrada equivalentes aos da criança;
iv. Tempo: deve aprender no mesmo tempo que uma criança típica faria;
v. Desenvolvimental: no processo de aprendizagem deve cometer os mesmos erros de uma criança;
vi. Cognitiva: com os mesmos recursos cognitivos que uma criança tem à sua disposição.

Embora nem sempre triviais de serem avaliados em situações concretas e específicas, tais critérios fornecem direções interessantes nas quais tanto sua avaliação quanto seu desenvolvimento podem seguir.

4. CATEGORIZAÇÃO DE PALAVRAS

O objetivo da pesquisa em andamento apresentada aqui é abordar um aspecto particular do processo de aquisição, a saber, a aprendizagem das categorias lexicais da língua. Por isso, é essencial que tenhamos bem definido o que são essas categorias e de que modo são identificadas. Segundo Basílio (2008), a questão dos critérios de classificação das palavras é muito debatida, por exemplo, quanto a se devemos usar um ou vários critérios, quais seriam melhores (se sintáticos, semânticos ou morfológicos).

Peters (1986) explica que, tradicionalmente, a descrição da fala adulta foi classificada pela distribuição de palavras baseada em seus papéis gramaticais, como, por exemplo, sujeito, verbo, objeto e modificador. Numa abordagem mais semântica, uma outra forma de classificação de palavras é pelos tipos de coisas que elas podem se referir, como, por exemplo, substantivos referem-se a entidades, verbos a ações e adjetivos a atributos. Um terceiro método de classificação é pelos papéis de caso, como: agente, ação, objetivo, benfeitor etc.

A descrição das palavras pode ser apresentada na forma de cada um desses métodos de classificação. Nenhum deles parece ser por si só suficiente, entretanto. Descrever gramaticalmente as classes de palavras deve ser algo feito simultaneamente por todos os critérios (BASÍLIO, 2008). A definição semântica, por exemplo, não inclui a posição de ocorrência das palavras na construção dos enunciados, o que é parte essencial da descrição

gramatical. A menos que seja possível deduzir o comportamento de uma certa classe a partir de sua função semântica, a definição por critérios semânticos não é suficientemente adequada para a descrição gramatical; como fazem as gramáticas escolares definindo, por exemplo, substantivos como palavras que designam seres.

Já a definição sintática mostra as posições estruturais de cada classe, mas não abrange outras propriedades que diferenciam as classes, como o substantivo ser classificado corretamente por sua posição como núcleo do sujeito, objetos e agente da passiva, mas não nas regras de concordância do substantivo com o adjetivo. Da mesma forma que só uma definição sintática e semântica não nos dá pistas suficientes de o porquê da necessidade de termos inúmeras formas verbais expressando categorias de tempo, modo, aspecto e número-pessoa. Por essas razões, é necessário que vários critérios sejam usados juntos (BASÍLIO, 2008).

Peters (1986) tenta entender como é feita na aquisição da primeira língua a categorização de palavras, não tendo a gramática adulta como ideal, mas sim, como a gramática infantil exibido suas próprias particularidades. A informação de que a criança dispõe inicialmente são de natureza semântica e distribucional por natureza, já que ela não tem ainda conhecimento sintático pleno. Assim, ela vai consolidar seu conhecimento linguístico formando classes de palavras. Por exemplo, as palavras que podem ocorrer com *more* podem ser agrupadas em uma classe distribucional, mas elas também irão compartilhar uma mesma propriedade semântica, que é a de serem o tipo de coisas recorrentes[6]. Maratsos (1982 *apud* PETERS, 1986) afirma ainda que mesmo que o conhecimento semântico seja uma base útil para formar categorias de classes de palavras nos primeiros estágios da aquisição da linguagem, não será suficiente a longo prazo.

Maratsos & Chalkley (1980 *apud* MINTZ ET AL., 2002) sugerem que as categorias gramaticais podem ser parcialmente aprendidas através de análises distribucionais do discurso como dado de entrada. Do mesmo modo, crianças podem seguir procedimentos similares para aprender sua língua nativa, visto que o aprendiz, nos estágios iniciais da aquisição da linguagem, deve inferir as categorias sintáticas sem ter inicialmente regras e restrições dos conhecimentos de gramática. Conforme Redington et al. (1998), é plausível supor informações distribucionais complementariam outras fontes (semântica, fonologia, prosódia e conhecimento inato).

Maratsos & Chalkley (1980 *apud* REDINGTON ET AL., 1998) notaram, por exemplo, que no inglês as palavras cuja raiz pode se combinar ao sufixo -ed e também ao sufixo -s são verbos (p.e., “photograph”, “like” etc.). Assim como palavras com o sufixo -s, mas não com o sufixo -ed, tendem a ser substantivos contáveis, (p.e., “kiss”, “bus”

6. Traduzido do original: “For instance, all the words that can occur with more can be grouped in a distributional class, but they will also share the semantic property of being the kinds of things that can recur”.

etc). Neste sentido, sugerem que o fato de que palavras de uma mesma categoria tendem a exibir regularidades distribucionais faz com que a evidência distribucional seja uma fonte de informação potencialmente importante para identificar categorias sintáticas das palavras, o que não exclui as demais fontes.

Kelly (1992 *apud* REDINGTON ET AL., 1998) propõe, por exemplo, que regularidades fonológicas acabam dando pistas sobre as categorias sintáticas, assim como a duração das palavras e os contornos prosódicos (p.e. variação rítmica e modulação no tom da voz). O autor dá o exemplo das palavras no inglês que são polissílabas, estas normalmente também são substantivas. Van Heugten et al. (2014) ainda afirma que as palavras funcionais (aquelas com pouco sentido lexical: determinantes, auxiliares e pronomes) tendem a ter, além de uma duração menor, uma força fonológica inferior e menos intensidade na produção, se comparadas às palavras de conteúdo (aquelas opostas às funcionais). Segundo as autoras, regularidades fonológicas podem ajudar a criança a refinar as unidades sintagmáticas e a detectar a ordem básica das palavras dentro dos sintagmas, uma vez que as frases entoacionais tendem a ser universalmente marcadas por pistas fonológicas, tais como a força do começo da frase, o alongamento do final das frases, o declínio do tom e as pausas, por exemplo.

5. APRENDIZAGEM DISTRIBUCIONAL

Visto que o principal estudo no qual esta pesquisa está baseada é o método distribucional apresentado em Redington et al. (1998), ele deve ser cuidadosamente discutido. Pois os autores demonstraram empiricamente que informações distribucionais fornecem uma poderosa pista para categorias sintáticas associadas, que podem ser exploradas por uma variedade de mecanismos simples e psicologicamente plausíveis.

Harris (1954) introduz o conceito de distribuição como: "a distribuição de um elemento será entendido como a soma de todos os contextos em que ocorre" (p. 146). O autor ainda lista quatro itens que ajudam a confirmar que a estrutura distribucional está presente na língua: (i) os elementos não aparecem randomicamente; (ii) os elementos de uma classe têm uma mesma frequência relativa, porém elementos de classes diferentes não possuem tal semelhança; (iii) podemos ter expectativas das estruturas que serão faladas a partir de um momento que o falante enuncia algo, mas é impossível prever o conteúdo do que será dito; e (iv) as restrições das categorias vão ficando cada vez mais específicas à medida que se categoriza mais as classes. É preciso atentarmo-nos ainda para duas questões: a primeira é de que, para o autor, a estrutura realmente existe na linguagem e a segunda é de que a estrutura realmente existe nos falantes, assim como defende a Gramática Gerativa. Ainda segundo Harris, é evidente que certos comportamentos dos falantes indicam uma percepção que concorda com os argumentos da estrutura

distribucional. A formação de novos enunciados na linguagem seria, portanto, baseado nas relações distribucionais.

Redington et al. (1998) comentam duas visões críticas ao método distribucional: a primeira o rejeita em princípio, pois isso seria transpor uma metodologia da linguística tradicional para o âmbito cognitivo da criança. Esse argumento falha, porém, ao não reconhecer a diferença entre a natureza das informações distribucionais usadas pelos linguistas e a informação distribucional que estaria disponível para a criança. A segunda, defendida por Pinker (1984), incide nas limitações do método e na não-universalidade das propriedades distribucionais observadas nas línguas.

No entanto, segundo Redington et al. (1998), argumentam que avaliar a abordagem distribucional com base em estudos simplistas e instáveis não é um argumento válido. Ademais, a solução ao debate é no fim uma questão empírica, o que justifica estudos como o apresentado aqui. Ainda segundo os mesmos autores, este é um problema que parece particularmente tratável computacionalmente, assim como também ressalta Yang (2011), o que torna sua investigação ainda mais atrativa. A seguir, apresentamos o estudo de Redington et al. (1998). No entanto, outros estudos similares foram feitos e também servem como referência para esta pesquisa (MINTZ ET AL., 2002; entre outros).

5.1 O estudo em Redington et al. (1998)

Para Redington et al. (1998), a informação distribucional refere-se a informações sobre contextos linguísticos em que uma palavra ocorre. O fato de palavras de uma mesma categoria tenderem a ter um alto número de regularidades distribucionais em comum faz com que as regularidades possam ser usadas como pistas para as categorias sintáticas. Os autores sugerem ainda que os métodos distribucionais não sejam propostos como uma solução geral para o problema do aprendizado da linguagem, mas sim como uma potencial fonte de informação sobre as estruturas sintáticas. Além do mais, é provável que haja restrições inatas na possível análise distribucional e nos mecanismos de aprendizagem que o aprendiz aplica, e é razoável, porém não necessário, que esses mecanismos de aprendizagem sejam específicos para a linguagem. Dessa forma, métodos distribucionais poderiam eles mesmos, de uma certa forma, incorporar o conhecimento inato.

Para testar o método distribucional e demonstrar que as propriedades distribucionais das palavras podem ser muito informativas sobre uma categoria sintática – e que tais propriedades podem ser extraídas por mecanismos psicologicamente plausíveis –, Redington, Chater e Finch desenvolveram nove experimentos com base em um corpus de fala dirigida à criança a partir de dados da base CHILDES. Após remoção de material considerado irrelevante, o corpus consistiu em 2,5 milhões de palavras. Não houve normalização da ortografia, de modo que ocorrências como *wanna* e *wannaa*

foram consideradas palavras distintas. Para avaliação das classificações, foi feita uma categorização de referência desse corpus usando um banco de dados contendo frequências de categorias sintáticas de palavras. Cada palavra foi categorizada com a categoria sintática mais frequente (a modelagem não lida com ambiguidade). Para a classificação padrão, foram usadas as 1.000 palavras mais frequentes como palavras-alvo, as 150 mais frequentes como palavras contextuais e os enunciados foram concatenados como uma longa cadeia ininterrupta.

A análise padrão para os experimentos teve a seguinte característica: corpora de fala direcionada à criança com 2,5 milhões de palavras, as 1000 palavras-alvo mais frequentes, as 150 palavras de contexto mais frequentes, e duas palavras contextuais de cada lado. Nos experimentos específicos, foram comparados seus resultados com os da análise padrão e com uma linha aleatória, sendo as principais medidas, as de acurácia e completude. Além das variáveis próprias a cada experimento, foi preciso repetir cada experimento variando o nível de similaridade assumido (0.1 a 0.9) e gerou-se um *baseline* também nessas condições (além do *baseline* da condição).

O método dos autores envolve três etapas: (i) medir a distribuição de contextos em que cada palavra ocorre; (ii) comparar as distribuições dos contextos para pares de palavras; e (iii) agrupar palavras com distribuições de contextos similares. No caso de (i), coletam-se estatísticas de co-ocorrência entre a palavra alvo e as palavras em seu entorno, o que pode ser representado como um *vetor contextual*. Dentre as várias medidas de similaridade possíveis entre dois vetores contextuais, os autores optam pelo *coeficiente de correlação de postos de Spearman* (ρ) (ver Redington et al., op.cit., para uma discussão sobre essa medida). Para agrupar palavras em grupos, os autores usam o método "standard hierarchical cluster analysis" (SOKAL & SNEATH, 1963 *apud* REDINGTON ET AL., 1998). O algoritmo começa por agrupar itens próximos de acordo com uma métrica de similaridade. Assim que um grupo é formado ele pode ser agrupado com outros itens ou outros grupos. A distância entre dois grupos é a média das distâncias entre os membros de cada um.

Abaixo, são apresentados os nove experimentos concebidos para avaliar diferentes parametrizações ou aspectos do problema. É apresentado um breve sumário de cada um, incluindo resultados obtidos:

1. *Diferentes contextos*: avaliou a informatividade de cada uma das quatro palavras que precedem a palavra-alvo, assim como cada uma das quatro palavras que a sucedem. Avaliou-se também a combinação de posições, tais como uma anterior e uma posterior, ou duas anteriores e duas posteriores. O experimento 1 mostrou que contextos mais locais são os mais informativos sobre as categorias sintáticas.

2. Variação do número de palavras-alvo e contextuais: neste experimento variou-se o número de palavras usadas como alvo (entre 100 e 2000 mil) e como contextuais (entre 10 e 500), mantendo todos os outros aspectos de análise constantes. O método apresentou uma curva em U invertida, sendo os seus picos em 1000 palavras-alvo e 150 palavras de contexto. Os autores afirmam por fim que, de modo geral, o método continuou sendo mais ou menos informativo em todas condições.
3. *Para quais classes a informação distribucional é valiosa*: avaliou a utilidade das informações distribucionais para diferentes classes de categorias sintáticas obtidas pela análise e se isso refletia propriedades da aquisição infantil de certas categorias. Verificou-se que a análise distribucional foi mais informativa sobre a classe dos substantivos. A performance para os verbos também foi considerável, porém menos do que a dos substantivos. A performance dos adjetivos foi moderadamente boa, mas a dos advérbios foi relativamente pobre. Além disso, os resultados obtidos para as palavras de conteúdo foram melhores do que para as funcionais. Esses resultados, segundo os autores, são compatíveis com dados desenvolvimentais.
4. Tamanho do corpus: investigou se as análises poderiam ser efetivas ao usar a quantidade de dados de entrada normalmente disponível para as crianças. Segundo os autores, haveria em torno de 1,5 milhões de palavras na fala dirigida à criança em um ano. Para saber a quantidade de dados de entrada necessária para fornecer informações úteis, foi avaliada a eficácia do método para diferentes tamanhos do corpus (100 mil,500 mil, 1 milhão e 2 milhões de palavras). Resultados: a aprendizagem distribucional de categorias sintáticas se mostrou melhor que a baseline em todas as análises, embora foi a partir de 1 milhão que foram significativamente melhores.
5. *Fronteiras de enunciado*: avaliou se o conhecimento das fronteiras de enunciado poderia fornecer informações adicionais sobre as categorias sintáticas. Assim, as fronteiras foram delimitadas de duas formas: (i) as palavras de contexto fora do enunciado não foram contabilizadas (isto é, fronteira nula); (ii) tratar as fronteiras de enunciado como um item lexical explícito a ser contabilizado. Os resultados demonstram que a fronteira explícita é mais informativa para a classificação das categorias, o que indica que o falante explora essa informação.
6. *Frequência x ocorrência*: investigou a eficácia do método usando apenas uma contagem binária da ocorrência de um dado item como contexto, ao invés de sua frequência. Os resultados mostram que, embora a contagem binária dê resultados ligeiramente acima da *baseline*, o uso da frequência é o que traz a maior informatividade.
7. *Removendo palavras funcionais*: testou a possibilidade de as crianças prestarem mais atenção às palavras de conteúdo do que as funcionais. Qual seria a eficácia do aprendizado distribucional se palavras funcionais fossem removidas dos corpora?

A ideia é simular uma possível insensibilidade da criança aos itens funcionais nos primeiros estágios. A remoção dos itens funcionais impactou negativamente a performance do método, embora ainda acima da *baseline*, e, portanto, útil ao aprendiz.

8. *Informação de uma categoria ajuda a aquisição de outras?* Como visto anteriormente, a criança é suscetível a explorar as variedades de pistas sobre a linguagem, e estas podem ser usadas em combinação com a análise distribucional. Dessa forma, é interessante se perguntar qual a dimensão das dicas das outras fontes de informação sobre as classes dos itens que podem ajudar a análise distribucional. Para isso, avaliou-se se a substituição de palavras de uma categoria pelo nome da categoria (p.e., verbos específicos por VERB) afeta a categorização. De modo geral, as análises mostraram uma diminuição da informatividade sempre que itens lexicais específicos foram substituídos pelas suas categorias. Isso sugere que a frequência de itens lexicais é mais informativa do que as categorias das palavras já conhecidas.
9. *A aprendizagem é mais fácil com fala dirigida à criança*? A questão é se a língua dirigida para a criança facilita a aquisição de categorias sintáticas, isto é, se o "manhês" de fato contribui para a aquisição. Este experimento, então, verificou se há diferenças qualitativas entre a fala dirigida à criança e a fala entre adultos no que diz respeito à informação distribucional. As análises mostraram que há pouca diferença qualitativa entre as falas, com uma leve vantagem para a fala entre adultos, o que sugere que o manhês não facilita a vida da criança, pelo menos não nos aspectos distribucionais. No entanto, é imprescindível notar que o corpus de fala dirigido à criança usado por Redington et al. (1998) continha também fala entre adultos, o que pode ter influenciado os resultados.

Redington et al. (1998) concluem que seus resultados demonstram que a informação distribucional é uma pista potencialmente poderosa para a aprendizagem de categorias sintáticas. Os autores ressaltam que, até o momento de seu estudo, não havia demonstrações computacionais similares para outras fontes de informação (fonológica, prosódica ou semântica). No entanto, hoje temos, por exemplo, estudos como os de Adriaans & Swingley (2012) e de Monaghan et al. (2007), que tratam sobre prosódia e fonologia, respectivamente.

Adriaans & Swingley (op.cit) examinam se as modificações na prosódia para a fala direcionada à criança auxilia o aprendizado distribucional das vogais. Já em Monaghan et al. (op.cit.), há apontamentos sobre a interação entre pistas fonológicas e distribucionais em quatro línguas diferentes (a saber: inglês, alemão, francês e japonês). Os autores indicam que, ao passo que as pistas distribucionais ficam menos confiáveis, as fonológicas ficam mais fortes, indicando que a linguagem é estruturada de tal modo que a aprendizagem

se beneficia da integração de informações (desse modo, o ambiente da criança torna-se menos pobre do que imaginávamos). Redington et al. (1998) ainda ressaltam que os estudos translinguísticos são necessários para avaliar mais profundamente o papel da informação distribucional. Ressaltamos, assim, a importância do presente estudo no sentido de agregar ainda mais resultados aos já obtidos e contribuir para uma melhor compreensão destes aspectos da aquisição da linguagem.

6. O ESTADO ATUAL DA INVESTIGAÇÃO

A pesquisa apresentada aqui está em fase de andamento, portanto, não temos resultados experimentais. Até o presente momento, focamos em obter os corpora e prepará-los para os experimentos. Além disso elaboramos algoritmos-base para a implementação do método distribucional nos dados.

6.1 Corpora para análise

A análise do método distribucional será feita em cima dos corpora da base de dados CHILDES (Ver Figura 1) (MACWHINNEY, 1989), da Coleção "Projeto Aquisição da Linguagem Oral" (Ver Figura 2), disponível no CEDAE/IEL, e do Projeto NURC/RJ ("Projeto Norma Linguística Urbana Culta - RJ") (Ver Figura 3), sendo este último de fala entre adultos, enquanto os dois primeiros são de fala direcionada à criança. Essa preparação envolveu a obtenção dos dados nas sua respectivas fontes. O corpus do NURC foi acessado online e cada transcrição foi salva em formato texto (.txt). Para o corpus do Projeto de Aquisição de Linguagem Oral foram feitas visitas ao CEDAE nas quais tivemos acesso à porção do acervo transcrita em formato PDF, a partir dos quais geramos versões em formato texto. O corpus do CHILDES ainda está em processo de preparação.

O corpus do Projeto de Aquisição de Linguagem Oral preparado é composto de 411 arquivos de 8 crianças diferentes, com cerca de 753 mil palavras. Já o corpus do NURC consiste de 167 arquivos, com um pouco mais de 1,1 milhão palavras. Ainda não temos os dados sobre o corpus CHILDES.

Assim como no estudo de Redington et al. (1998), optamos por não normalizar as transcrições. Portanto, podem haver desvios de ortografia, como ocorrem algumas vezes nos verbos no infinitivo (principalmente nas falas direcionadas à criança), que, ao invés de estarem como "cantar", estão "cantá", por exemplo.

*MAE: uma de sopa.
*MAE: queres carne?
*MAE: +< qué.

*ALS: ela disse é@q ?
*MAE: qué@q .
*ALS: ah qué@q .

Figura 1: trecho de diálogo entre a mãe, o investigador e a criança (fala oculta)

Ad.: Canta pra mim uma musiquinha, pra ver se você sabe cantá. Ad.: Não… Na escolinha você canta… Não canta?

Ad.: Sabe sim. Você cantou.

Ad.: Não. Botãozinho não pode apertar não. O que aconteceu?

Ad.: Pronto. Mais uma vez e chega. Não põe a mão não, dá choque. Ad.: Anamaria, o que você faz na escolinha hoje?

Ad.: Teve aniversário hoje na escolinha?

Figura 2: trecho de diálogo entre um adulto e a criança (fala oculta)

LOC.: Bom, lanchonetes, né? A salvação para o verão, né? Refrigerantes, água mineral, embora ache que o que mate mesmo a sede é a água pura, potável. Agora os bares, para os, naturalmente os maiores de idade, não é, são muito falsificados também, acho que deviam ter mais controle. As bebidas estrangeiras são falsificadíssimas e são, os preços, explorados. DOC.: O que que as pessoas bebem nos bares?

LOC.: Depende do, da classe, né? As classes mais altas uísque, acho que a grande maioria, né, gim, rum, e a classe mais baixa fica na cachaça mesmo, na batida, pra tapear, né, cachaça disfarçada, né?

DOC.: E os mais jovens

Figura 3: trecho de diálogo entre dois adultos

6.2 Algoritmos para o método distribucional

O algoritmo apresentado abaixo, feito inicialmente para processar os dados do corpus NURC, tem a função de (a) carregar os dados do arquivo; (b) remover os metadados no início do arquivo; (c) separar as falas em ID do falante e FALA; (d) tokenizar a FALA usando a NLTK; e (e) retornar a lista de pares (ID, FALA). Assim, temos toda a amostra do NURC organizada como uma lista de pares indicando o falante e seu enunciado. Posteriormente, o algoritmo será adaptado para processar os outros corpora.

```
função load_nurc() {

Obter a lista de arquivos;

dados := [];              // Lista de tuplas (falante, enunciado)

Para cada arquivo faça { Abrir o arquivo;
Ler o conteúdo do arquivo;
texto := Extrair dados dos falantes do conteúdo;

enunciados := separar_por_linha(texto);

para cada linha {
// Separar o ID do falante e seu enunciado, a partir do separador "." datum := separa_id_fala(linha);

adicionar_na_lista(dados, datum);
}

}
retornar dados;

}
```

Figura 4: código de tratamento de enunciados dos corpora do NURC

O segundo algoritmo é o primeiro passo do método distribucional em Redington et al. (1998), cujo objetivo é medir a distribuição de cada palavra. Para isso, precisamos: (a) coletar os contextos; (b) medir a coocorrência entre pares; (c) fazer uma tabela de contigência; (d) considerar poucas palavras-alvo; (e) restringir palavras de contexto; (f) obter vetor de contexto; (g) juntar vetores; e (h) ter a representação final .Por enquanto, temos um código que recebe o número de palavras de contexto das palavras-alvo que devem ser consideradas, tanto anteriores quanto posteriores, e é calculada a frequência em que elas ocorrem dentro da janela de ocorrência pré-estabelecida.

```
função calcula_contextos(

        texto,                                          // Texto a ser processado
        posicoes_de_contexto, // Posições contextuais a considerar palavras_de_contexto, // Qtde
        de palavras de contexto palavras_alvo)     // Qtde de palavras-alvo

{

    tabelas := Criar_tabelas_de_contingência(posições_de_contexto); lista_contexto := Obter_
    mais_frequentes(texto, palavras_de_contexto); lista_alvos := Obter_mais_frequentes(texto,
    palavras_alvo);

    Para i de 0 até tamanho(texto), faça {
    palavra = texto[i];
        Se palavra está em lista_alvos, então { Para cada tabela de contingência, faça{
    contexto := obter_palavra(texto, i, posição_contextual); Se contexto está em lista_contexto, então{
```

```
        incrementar_contagem_na_tabela(palavra, contexto);
              }
            }
          }
        }
retorna tabelas;
}
```

Figura 5: código do primeiro passo do método distribucional de Redington et al. (1998)

7. CONCLUSÃO

O modelo proposto por Redington et al. (1998) de como as informações distribucionais são informativas para as crianças adquirirem as categorias sintáticas sugere que estas são potencialmente importantes no processo de aquisição no inglês. A pesquisa aqui apresentada visa estender estes resultados para o português brasileiro, o que nos coloca o desafio de compreendermos métodos e cálculos estatísticos e matemáticos, além da criação dos algoritmos que lidem com todas essas variáveis. Essas etapas ainda estão em fase de desenvolvimento na presente pesquisa. Trabalharemos, assim, para que consigamos replicar o experimento de Redington et al. (1998) com os dados do português brasileiro, analisando e comparando os resultados dos dois estudos, a fim de que seja uma pesquisa mais fidedigna possível, a fim de que tenhamos resultados bons que contribuam para a área da aquisição.

REFERÊNCIAS

Adriaans, F., & Swingley, D. (2012). Distributional learning of vowel categories is supported by prosody in infant-directed speech. *Proceedings of the Cognitive Science Society*, 34(34),22-77.

Basilio, M. (2008). Classe de palavras e categorias lexicais. *Formação e classes de palavras no português Brasil*. Editora Contexto.

Costa, J., & Santos, A. L. (2003). A gramática que os bebés sabem. *A falar como os bebés: o desenvolvimento linguístico das crianças*.

Faria, P. (2013). Um modelo computacional de aquisição de primeira língua. Tese de doutorado. Unicamp.

Grolla, E., & Figueiredo Silva, M. C. (2014). A capacidade linguística de adultos e crianças. Para conhecer: aquisição da linguagem. *São Paulo: Contexto*.

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

Ingram, D. (1989). *First language acquisition: Method, description and explanation*. Cambridge university

press.

Kaplan, F., Oudeyer, P. Y., & Bergen, B. (2008). Computational models in the debate over language learnability. *Infant and Child Development*, 17(1), 55-80.

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

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26(4), 393-424

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive psychology*, 55(4), 259-305.

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

Peters, A. M. (1986). Early syntax. *Language acquisition*, 2, 307-325.

Pinker, S. (1984). Language learnability and language learning. *Cambridge, MA: Harvard*.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive science*, 22(4), 425-469.

Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental psychology*, 22(4), 562.

Van Heugten, M.; Dautiche, I.; Christophe, A. (2014). Phonological and prosodic bootstrapping. *ENcyclopedia of Language Development*, 447-451.

Yang, C. (2011). Computational models of syntactic acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 3*(2), 205-213.