

Desafios metodológicos à Linguística de Corpus em aquisição da linguagem no português brasileiro

Methodological challenges to Corpus Linguistics in language acquisition in Brazilian Portuguese

Leonor Scliar-Cabral

Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina, Brasil

Resumo: Discutirei os desafios metodológicos para a organização de três *corpora* de uma criança aos 20 meses e 21 dias, 22 meses e 20 dias e 26 meses e 8 dias, ao adquirir o PB. Utilizo a metodologia de Brown (1973) quando levantou os *corpora* de Adam, Eve e Sarah, empregando o cálculo de Extensão Média de Enunciados, EME (MLU) e de Limite Máximo de Enunciados (*Upper Bound*), cujas regras são descritas. Os resultados são confrontados com dezenove relatos de Brown (1973, p. 66) referentes à aquisição de línguas diferentes, inclusive, não indo-europeias. Explicam-se os critérios para seleção dos enunciados dos *corpora* para depreender as gramáticas emergentes. O total dos enunciados da criança em transcrição fonética foi de 5530, o dos adultos, transcritos canonicamente, quando dirigidos à criança, foi de 5664. As gravações, transcrições fonéticas, glosas, análise morfológica (linha *mor*) e léxicos, estão disponíveis na Plataforma CHILDES, o maior banco mundial de dados sobre aquisição da linguagem. Abordo os desafios metodológicos, inclusos os enfrentados pelos pesquisadores que trabalham com aquisição do PB, advindos das novas tecnologias, em especial, as assumidas pela Plataforma CHILDES que adota na linha Mor (análise morfossintática) as mesmas medidas explicadas no artigo.

Palavras-chave: Linguística de *corpus*; Aquisição do PB; Desafios metodológicos; Extensão Média de Enunciados

Abstract: I will discuss the methodological challenges for a child aged 20 months and 21 days, aged 22 months and 20 days and aged 26 months and 8 days three corpora organization, when acquiring Brazilian Portuguese. I used Brown's (1973) methodology when he raised Adam, Eve and Sarah's corpora using Mean Length Utterance, MLU and Maximum Upper Bound Limit calculation, whose rules I describe. Results are compared with the nineteen reports by Brown (1973, p. 66) regarding different languages acquisition, including non-Indo-European ones. Criteria for selecting utterances for making up emerging grammars are also explained. Child's total utterances number, in phonetic transcription, was 5530, while the adults' number ones, canonically transcribed, when addressed to the child, was 5664. All



recordings, as well as phonetic transcripts, glosses, morphological analysis (mor line) and lexicons are available on the CHILDES Platform, the world's largest database on language acquisition. I address the methodological challenges, including those faced by researchers working with PB acquisition, arising from new technologies, in particular, those assumed by the CHILDES Platform that adopts the Mor line (morphosyntactic analysis) using the same measures here explained.

Keywords: Corpus Linguistics; PB Acquisition; Methodological Challenges; Mean Length Utterances

1 Introdução

Discutirei os desafios metodológicos para a organização dos três *corpora* de uma criança ao adquirir o português brasileiro, doravante PB, variedade sociolinguística paulista dos pais com nível universitário, em três cortes transversais: aos 20 meses e 21 dias, aos 22 meses e 20 dias e aos 26 meses e 8 dias. Segue-se a metodologia de Brown (1973, p. 51-59), quando levantou os clássicos *corpora* de Adam, Eve e Sarah.

Realizei a pesquisa no apartamento, presentes também a mãe, muitas vezes o pai e, excepcionalmente, a empregada, em situações de brinquedo, assistindo a vídeo, alimentação e higiene, sem visibilidade do gravador.

O primeiro levantamento durou cinco horas: foram gravados 1319 enunciados da criança, transcritos foneticamente com os padrões de entoação. Aqui, surge o primeiro desafio. Conforme Brown (1973, p. 53, tradução minha), “Qualquer um que planeje empreender um registro fonético necessita ser advertido de que se trata de um trabalho imenso”. Há que acrescer que os enunciados iniciais contêm sons, muitas vezes, distantes dos que realizam os fonemas do PB, além de estruturas silábicas inexistentes, como [mbe], fusão da frase ‘um beijo’. As gravações e transcrições fonéticas estão disponíveis na Plataforma CHILDES, o maior banco mundial de dados sobre aquisição da linguagem. No primeiro levantamento gravei mais 1313 enunciados dos adultos dirigidos à criança e outros tantos não numerados, falados entre si e transcritos canonicamente.

O segundo levantamento durou seis horas: gravei 2245 enunciados da criança, transcritos foneticamente, mais 2475 enunciados dos adultos dirigidos à criança e outros tantos não numerados, falados entre si e transcritos canonicamente. O terceiro

levantamento durou seis horas: gravei 1996 enunciados da criança e 1876 dos adultos dirigidos à criança e outros tantos falados entre si.

O total dos enunciados da criança em transcrição fonética foi de 5530. O dos adultos, transcritos canonicamente, dirigidos à criança, foi de 5664.

A pesquisa longitudinal foi em situação natural, tanto quanto possível. Orientei os pais para manterem comportamento espontâneo sem forçar a criança a enunciados estereotipados. Durante a gravação, tomei notas sobre a situação ambiental e fiz uma primeira transcrição grosseira, pois, na gravação, as sílabas átonas finais de palavra carecem de energia e se tornam inaudíveis.

As gravações, realizadas em gravador Sony, *stereo cassette-corder*, modelo TC-146^a, iniciaram em 1974. As fitas utilizadas na fase 1 são Sony Auto-Sensor C-60 HP; as das fases 2 e 3, Sony Auto-Sensor C-90 HP.

2 Procedimento de transcrição

Realizei uma primeira transcrição manuscrita e uma segunda, datilografada, com revisão cuidadosa das fitas e checagem da primeira transcrição, por duas alunas do Curso de Fonoaudiologia da então Escola Paulista de Medicina, hoje UNIFESP, fundada em 1994.

Usei o IPA para a transcrição lata e nos padrões de entoação, usei o sistema de Matluck (1965). Sob cada transcrição fonética há uma glosa com interpretação em transcrição canônica. A numeração à esquerda da página indica os pontos do conta-giro, de 50 em 50, para controlar a fita e localizar rapidamente qualquer enunciado. Assinalei os nomes dos adultos pela inicial para mantê-los anônimos. A seguir, com o respectivo número, o enunciado dirigido à criança, em transcrição canônica. Incluí os enunciados dos adultos quando falavam entre si, sem numeração, para eventuais pesquisas para contrastar com o registro utilizado com a criança e possíveis influências desse último, na aquisição do léxico. Os enunciados da criança vêm do lado direito, precedidos da numeração. Bem à direita, consta o número de itens por enunciado, para cálculo da Extensão Média, EME (no inglês, *Mean Length Utterance, MLU*).

No texto, utilizei o nome Pá, referente à criança, um apelido, sem perigo de identificação. Os pais da criança, ele, linguista, ela, psicóloga realizaram um confronto entre a transcrição datilografada e as fitas. Finalmente, fiz nova checagem.

A seguir, segmentei os enunciados e organizei o léxico, pelos modelos da gramática gerativo-transformacional (CHOMSKY, 1964) e da Gramática dos Casos de Fillmore (1968, p. 1-18) e, para fins de cálculo da EME e do Limite Máximo de Enunciados (*Upper Bound*), segui os critérios de Brown (1973, p. 54), traduzidos por mim.

3 Regras para cálculo da Extensão Média de Enunciados e do Limite Máximo de Enunciados

- a. Começar pela segunda página da transcrição, salvo se incluir recitativo. Nesse caso, começar pelo primeiro trecho sem eles. Contar os primeiros 100 enunciados que satisfaçam as condições;
- b. Utilizar só enunciados completamente transcritos e nenhum com espaços em branco. Usar trechos entre parênteses para indicar transcrição dúbia;
- c. Incluir todas as repetições exatas de enunciados (assinaladas com o sinal + nas transcrições). Assinalar os titubeios como esforços repetidos de uma só palavra: contar como uma só ocorrência a forma produzida mais completamente. Nos poucos casos em que uma palavra for pronunciada mais enfaticamente (não, não, não), computar cada ocorrência;
- d. Não computar pausas plenas como ‘mm’, ‘ə’ (adaptação ao PB do último exemplo), mas, sim, ‘não’, ‘é’, ‘oi’ (adaptação ao PB dos três exemplos);
- e. Computar como uma só palavra todas as palavras compostas (de dois, ou três morfemas livres), nomes próprios e reduplicações ritualizadas. Por exemplo: mbê (um beijo); pique-pique-pique (onomatopeia imitativa das picadas da agulha), exemplos adaptados ao português, por não haver evidência de que os morfemas constitutivos funcionem como tal para a criança;
- f. Contar como um só morfema todos os passados dos verbos irregulares (cf. ‘fez’), pois não existe evidência de que a criança os relacione às formas do presente;
- g. Contar como um só morfema os diminutivos (no PB, ocorrem também os aumentativos, formas básicas, conforme ‘Paião’, ‘grandão’) porque a criança pesquisada não usa os sufixos de modo produtivo;

- h. Contar como morfemas separados todos os auxiliares ('é', 'tá', 'qué', 'vai', 'pode'). Igualmente, os catenativos, como 'é-pra', 'tem-de', esses últimos, computados como um só morfema, pois assim funcionam para a criança. Contar como morfemas separados todas as flexões, por exemplo, o plural -s; o morfema de tempo-aspecto-pessoa -w do pretérito perfeito do indicativo, ou de gênero -u, -a;

Figura 1 - Primeira página do *corpus* de Pá, aos 20 meses e 21 dias.

CORPUS, Fita 1, lado A	
Sujeito: Pá, Idade: 20m, 21d Data da gravação: 16/03/1974	Participantes I: pai E: mãe L: pesquisadora
CF ADULTO	CRIANÇA (brincando com a máquina de escrever velha, no gabinete, sentado no chão)
000 I: 1 De quem é a maquiqui? 2 De quem é a maquiqui? 3 É meu, não é? 4 É meu?	3 ↑ 1 m
5 É meu a maquiqui, né? 6 Então eu vou levar.	3 1 ↓ ~ ~ 2 'nə :w ↓ não!
7 Não? Por quê?	2 ↓ ~ ~ 3 'nəw ↓ não!
8 É meu!	2 ↓ ~ ~ 4 'nəw → não!
E: 9 De quem é? I: 10 De quem é? 11 De quem é a maquiqui?	2 ↓ ~ ~ 5 'nəw → não!
12 Vamos brincar um pouco, hoje? L: -Vocês ignorem a minha presença. E: 13 Quer um papel, nenê, quer?	2 1 ↓ ~ ~ 6 'nəw 'ε ↓ não é
14 Quer um papel? 15 Quer? 16 A mamãe vai ficar trabalhando aqui. 17 Você fica aí, tá? 18 Olhe, P.! 19 Olhe! 20 P.! 21 P.!! 22 P.!!! (assoviando para chamar a atenção)	2 3 1 ↓ 7 'ε 'εpe ↓ é minha (?)
	3 1 ↓ 8 'εpe ↓ é minha (?)
	2 1 ↓ 9 'εpe ↓ é minha (?)
	1 ↓ 10 'k ^h aw ↓ carro (?)
	3 ↑ 11 m
	3 ↑ 12 m
	Nº
	1
	1
	1
	1
	1
	2
	2 (?)
	2 (?)
	2 (?)
	1
	1

Fonte: Scliar-Cabral (1976b).

A contagem segue as regras acima, mas calculada para o total das transcrições e não apenas para os 100 enunciados iniciais.

Escolhi os critérios de Brown para cálculo da EME, em primeiro lugar, para confrontar os resultados obtidos na aquisição do PB com os de pesquisas em outras línguas, os dezenove relatos de Brown (1973, p. 66), que utilizaram os mesmos critérios.

Quanto mais semelhantes as características encontradas na aquisição de línguas pertencentes a famílias linguísticas bem distintas, tanto mais se mostra correta a hipótese de universais linguísticos na aquisição da linguagem.

Em segundo lugar, embora a EME apresente problemas que discutirei, tem se mostrado uma medida externa mais adequada que o dado cronológico. Existe uma correlação bem maior entre o EME e os aspectos característicos da maturação linguística, denominada lei de complexidade crescente por Brown, do que a verificada entre a idade cronológica e os aspectos acima referidos.

Em terceiro lugar, tal medida possibilita confrontar os resultados entre crianças adquirindo a mesma língua.

Mas tal medida ainda apresenta problemas, entre eles, o de muitos critérios terem sido estabelecidos *ad hoc* para o inglês, carecendo de generalização.

Em virtude de na fase 1, início, a gramática da criança apresentar-se com aparência pivotal, com a utilização de operadores aos quais ela adjunge itens em aberto, é necessária muita cautela, para não descartar itens com aparência de muletas, mero recurso articulatório de desempenho, na verdade, itens gramaticais produtivos. Os critérios para segmentar ou não os enunciados, explicitados por Brown, revestem-se, às vezes, de tom impressionista, carecendo de objetividade, sem especificação do que considera a evidência de que determinados morfemas, como os de passado, nos verbos, sejam improdutivos.

Na criança adquirindo o PB, usei os seguintes critérios para segmentar:

- 1) O item deve ter aparecido isolado e / ou combinado com mais de um morfema, apresentando, pois, distribuição que evidencia sua existência como unidade;
- 2) a significação deve recorrer com adequação nos contextos linguísticos e situacionais;
- 3) itens com a mesma significação recorrente e distribuição gramatical são variantes de uma mesma forma.

4 Critérios para escolher os enunciados e extensão da amostragem para deprender as gramáticas

Apliquei os seguintes critérios para selecionar os enunciados integrantes dos *corpora*, além dos mencionados nas disposições de Brown:

Para cada fase selecionei 713 enunciados para deprender as gramáticas, conforme a decisão da equipe de Harvard (BROWN, 1973, p. 56) e somente utilizei enunciados com mais de uma palavra, descartando, primeiro, aqueles com somente uma ocorrência quanto ao tipo de construção, por carecerem de produtividade, revelando, quase sempre, estágios incipientes de uma gramática, pois as ocorrências esporádicas tendem a se tornar produtivas nas fases subsequentes; segundo, os estereótipos por não revelarem a competência linguística da criança, como o auxiliar catenativo ‘é-pra’ (nº 99, 100, 444), sempre no mesmo contexto, seguido pelo verbo ‘pôr’, apresentando características de rotina e, terceiro, as imitações que requerem uma definição. Conforme Bowerman (1973, p. 21), considero imitação direta o retorno completo ou reduzido, na mesma ordem, do enunciado de outra pessoa, ocorrido até três enunciados anteriores.

Contudo, em confronto com as crianças comparadas por Bowerman (vide Tabela 2), o número de imitações de Pá é bastante pequeno. Tal diferença ocorreu porque considerei a entoação: se o adulto fizesse uma pergunta sim/não, com entoação ascendente, como “Você quer?” e a criança respondesse adequadamente com a entoação descendente “qué”, não o descartei, pois não o considerei imitação.

Tabela 1 - Dados ordenados evolutivamente.

Criança	Sexo	Eme	Idade do levantamento	Caráter do levantamento	Investigador	Língua
Eric I	M	1.10	1; 7	4 horas gravadas	Bloom	Inglês americano
Kendall I	F	1.10	-	2 dias completos, transcritos <i>in loco</i>	Bowerman	Inglês americano
Gia I	F	1.12	1; 7	7 horas gravadas	Bloom	Inglês americano
Eric II	M	1.19	1; 9	6 horas gravadas	Bloom	Inglês americano
Gregory	M	-	1; 7.5-1;11.5	Inventário cumulativo	Braine	Inglês americano
Andrew	M	-	1; 7.5-1;11.5	Inventário cumulativo	Braine	Inglês americano
Steven	M	-	1; 11.5-2;0.5	12 sessões de brinquedo, gravadas	Braine	Inglês americano
Christy	F	-	2; 0-2;3	Gravação semanal, sessões 45 min.	Miller, Ervin	Inglês americano

Susan	F	-	1; 8-2;0	Gravação semanal, sessões 45 min.	Miller, Ervin	Inglês americano
Kathryn I	F	1.32	1; 9	7 ½ horas gravadas	Bloom	Inglês americano
Gia II	F	1.34	1; 9	7 ½ horas gravadas	Bloom	Inglês americano
Eric III	M	1.42	1; 10	8 ½ horas gravadas	Bloom	Inglês americano
Seppo I	M	1.42	1; 11	2 horas gravadas durante 1 mês	Bowerman	Finlandês
Pá I	M	1.45	1; 8	5 horas gravadas em um 1 dia	xxxxxxxxxx	PB
Kendall II	F	1.48	1; 11	1 ½ horas, gravadas durante 2 dias	Bowerman	Inglês americano
Viveka	F	1.50	1; 11	4 ½ horas, gravadas em 1 mês	Rydin	Sueco
Sipili	M	1.52	2; 6	6 ½ horas, gravadas durante 1 semana	Kernan	Samoano
Tofi	M	1.60	2; 2	2 horas, gravadas durante 1 semana	Kernan	Samoano
Eve I	F	1.68	1; 6	3 ½ horas, gravadas durante 6 semanas	Fraser, Brown	Inglês americano
Sarah I	F	1.73	2; 3	3 horas, gravadas durante 6 semanas	Cazden, Brown	Inglês americano
Seppo II	M	1.81	2 ; 2	2 horas, gravadas durante 1 mês	Bowerman	Finlandês
Rina I	F	1.83	2; 1	2 horas, gravadas durante 1 mês	Bowerman	Finlandês
Pepe	M	1.85	2; 6	4 horas, gravadas em 2 dias consecutivos	Tolbert	Espanhol (México)
Kathryn II	F	1.92	1 ; 11	9 horas gravadas	Bloom	Inglês americano
Adam I	M	2.06	2 ; 3	2 horas, gravadas durante 1 mês	Bellugi, Brow	Inglês americano

^a Os algarismos romanos depois dos nomes das crianças foram estipulados pelos investigadores e se referem à ordenação de cada uma de suas análises; todas as crianças estão na fase 1, por definição. (Confronto com a tabela de Brown 1973, p. 66, tradução minha)

Fonte: Seliar-Cabral (1976a, p. 34).

5 Cálculos das medidas de Brown

5.1 Tipos de enunciado (utterance type)

Defino como tipo de enunciado o produzido com palavras isoladas, ou combinadas, não sendo imitação. Pá apresentou, na fase de EME 1.45, 274 tipos de enunciados, figurando entre Seppo e Sarah (vide Tabela 2).

5.2 Ocorrência de enunciados (utterance tokens)

Defino como ocorrência de enunciados todos os 713 enunciados da amostragem, repetidos, ou não.

5.3 Relação ocorrência-tipo (type-token ratio)

Defino a relação ocorrência-tipo como o quociente resultante da divisão dos tipos pelas ocorrências. Na fase de EME 1.45, a relação ocorrência-tipo foi de 0.38 (vide Tabela 2).

5.4 Limite máximo de itens por enunciado (upper bound)

A emergência das gramáticas infantis apresenta um aspecto dinâmico e não estático: numa dada fase, há uma gramática predominante (a que é descrita), mas persistem algumas regras e categorias da gramática precedente e já se anunciam as da subsequente: nessa última se enquadra o limite máximo de itens por enunciado.

Os pesquisadores se defrontam com soluções diversas para os limites impostos às gramáticas, quanto à capacidade de gerar cadeias mais extensas e complexas que configurem uma dada fase maturacional linguística. Assim, excluem as regras de recursividade, ou as que gerariam cadeias raras na criança. A literatura menciona duas soluções, a transformação de redução de Bloom (1970) e a do apagamento opcional do verbo de Bowerman (1973 p. 90 e segs.).

Tabela 2 - Características quantitativas das amostras dos sujeitos adquirentes do finlandês, do PB e do inglês.

	Seppo	Pá	Seppo	Rina	Adam	Eve	Sarah
EME	1.42	1.45	1.81	1.83	2.06	1.68	1.73
Idade (meses)	23	20; 21	26	25	27	18-19	27-28
Tipos de enunciado	297	274	437	338	505	309	265
Ocorrências de enunciado	713	713	713	713	713	713	713
Relação ocorrência/tipo	0.42	0.38	0.61	0.47	0.71	0.43	0.37
Limite máximo	5	5	5	5	7	4.25	4
Número de imitações	173	12	79	112	37	88	136
Porcentagem de imitações	24	1.68	12	15	4	12	18
Tamanho de léxico	112	145	226	136	201	169	193
Número de Substantivos	57	56	103	74	103	103	142
Número de verbos	31	30	65	32	37	36	21
Número de adjetivos	3	9*	8	5	15	6	13
Número de locativos	3	6	14	9	7	3	7
Número de pronomes	2	2**	9	8	7	5	?
Functivos requeridos	32	146***	98	168	127	91	81
Functivos presentes	1	46	8	16	7	14	15
Porcentagem de functivos presentes	3.1	31.5	8.2	9.5	6	13	16
(Comparação com a tabela de Bowerman (1973, p. 23), minha tradução)	Classes não mencionadas por Bowerman Fase 1, início (1.45):						

Cópulas	2
Auxiliares	3
Pronome possessivo	1
Partículas	3
Operadores ****	6
Demonstrativos	2
Interrogativos	2
Confirmativo	1
Tag	1
Interjeições	8
Onomatopeias	13
Estereótipos	1
*Incluídos os substantivos com função objetiva (possuidor) e os epítetos isolados	
**Pessoais	
***Levando em consideração os meus critérios, os functivos requeridos são 231 e os presentes, 48.	
**** ó; não, mais; outros, mais, um.	

Fonte: Seliar-Cabral (1976a, p. 38).

Adotei a solução de formular uma regra de condição impedindo à gramática gerar cadeias mais extensas. Contudo, o limite máximo de itens por enunciado é uma medida bastante reveladora, conforme a Tabela 2: em Pá, foi de 5.

5.5 Tamanho do léxico

O tamanho do léxico é obtido pela soma das palavras, com as respectivas repetições. Embora o número total por mim obtido não seja muito diferente das demais amostras, analisando item por item, verificar-se-ão algumas diferenças como o número de substantivos, porque realizei algumas discriminações: incluí substantivos isolados nas classes adverbiais (locativa, instrumental e companhia) e incluí a partícula interrogativa *m?* (123), com padrão ascendente que apresenta o maior número de ocorrências na amostragem, comprovando a modalidade interrogativa inequivocamente, conforme Menyuk (1971) e ao contrário de Bowerman (1973, p. 118) e Bloom (1973, p. 19)). Também incluí interjeições e onomatopeias, com articulação bem definida, empregadas com adequação semântica.

Contudo, os dados, como nas demais pesquisas, comprovam a incidência maior dos substantivos, seguida dos verbos e a quase total ausência de adjetivos, esses últimos, quase sempre, isoladamente, como epítetos: ‘feio’ (nº 185) e ‘sem-vergonha’ (nº 200), o que abona a hipótese de a causa primordial de não ocorrerem certas construções e classes sintáticas ser a complexidade linguística. Efetivamente, o adjetivo, embora contentivo, com carga semântica relativa ao mundo dos objetos e não à gramática e possuindo

saliência perceptual (todos possuem sílaba mais intensa), sem autonomia, atua como modificador do substantivo.

Ocorre apenas um tipo de modificação do substantivo, na Fase 1 de Pá, a possessiva, na construção N + N, que corresponde a uma categoria cognitiva já bem firmada e consignada por outros observadores, havendo apenas diferença quanto à ordem de colocação do possuidor e objeto possuído. Ocorre, também, a similaridade de o possuidor apresentar o traço [+animado] e o objeto possuído, os [-animado, +alienável].

Tais considerações reforçam a insistência de Brown para reformular o conceito de estilo telegráfico que comentarei a seguir.

6 Functivos requeridos (vide Tabela 2)

Os functivos (morfemas para Martinet (1964) e gramemas para Pottier (1968)) são as classes gramaticais caracterizadas por constituírem número limitado e fechado de elementos, pela alta frequência de ocorrência e por se referirem às significações internas da língua. No português, são functivos os artigos, preposições, conjunções, pronomes, advérbios, partículas (todas formas livres) e os afixos (desinências, prefixos e sufixos).

Característica da linguagem nas fases iniciais da aquisição é a ausência de functivos, daí a denominação de estilo telegráfico, pela semelhança com a linguagem dos telegramas.

Conforme Brown (1973, p. 74-90), no entanto, não se pode afirmar categoricamente a ausência dos functivos nas fases iniciais da aquisição e, sim, examinar quais variáveis influem para que alguns e não outros constem, que são: frequência alta *versus* baixa; ter independência silábica ou não; desempenhar um papel semântico ou apenas modular a significação de outro item; possuir carga informativa ou ser redundante; ser condicionado fonologicamente ou não; ser condicionado gramaticalmente ou não.

Brown conclui por que alguns functivos emergem antes que outros: “se o functivo apresentar frequência mínima, saliência perceptual alta, não for condicionado pelo contexto verbal, expressando um papel semântico básico, ao invés de uma modulação, então, será plenamente controlado na fase 1” (BROWN, 1973, p. 88)

O extremo oposto das variáveis mencionadas constitui condição para a ausência dos functivos. Na Fase 1, Pá não apresenta controle dos functivos, salvo daqueles que

atendam as variáveis de Brown, com exceção da 1ª (apresentar frequência mínima). Assim, ocorrem os prolocativos ‘aqui’ (‘aí’ é alomorfe em variação livre) e, ao contrário da maioria das pesquisas, as duas cópulas ‘é’, ‘tá’ e alguns auxiliares, que não preenchem todas as variáveis estipuladas, pois apresentam frequência muito alta. Ocorrem, ainda, alguns functivos como operadores, em virtude da aparência pivotal da gramática inicial.

Os autores, no cálculo dos functivos requeridos, seguem o critério de Cazden (1973) e somente computam lacunas em ocorrências de mais de uma palavra. Ora, nos enunciados de um só item, muitos deles acusam claramente a ausência de functivos; o mesmo ocorre com Bowerman (1973, p. 24-25) ao desconsiderar necessários os functivos quando tal necessidade depender mais da interpretação semântica atribuída ao enunciado do que da própria construção linguística. Por isso, modalizei os critérios com a norma de quais functivos são requeridos na fala usual, naquele contexto.

Assim, nos enunciados isolados 55 e 517 da Fase 1, depreende-se, indubitavelmente a ausência do functivo ‘na’ e, no enunciado isolado 68, a ausência de functivo da 1ª pessoa e, em geral, em todos os enunciados isolados dos verbos que se referem à 1ª pessoa do discurso, pois, essa lacuna comprova que a criança, na fase inicial, carece dos morfemas tanto presos quanto livres que assinalam a 1ª pessoa do discurso.

55. (na) mão

317. (na) rua

68. que (ro).

7 Desafios metodológicos com novas tecnologias

Comentarei os desafios metodológicos à linguística de *corpora* colhidos em pesquisas de aquisição da linguagem, com o advento de novas tecnologias que utilizam medidas semelhantes às recomendadas por Brown (1973, p. 51-59), na análise por amostragem da linguagem (Language Sample Analysis, LSA, RATNER; MACQUINNEY, 2016, p. 74), disponibilizadas no CHILDES.

Tais tecnologias são o KidEval acoplado ao CLAN e ao controlador Walker. O KidEval permite avaliar interações diádicas de uma vasta coorte de crianças e mães, desde os 7 até os 24 meses (NEWMAN; ROWE; BERNSTEIN RATNER, 2015), enquanto o CLAN, auxiliado pelo *playback* do controlador Walker, transcreve e analisa amostras da

linguagem (*op. cit.*, 2016, p. 76). Elas permitem economizar um tempo colossal requerido para as transcrições, numa proporção de três-quartos, além da maior acurácia.

Mas alguns dos desafios metodológicos persistem, exemplificados com a aquisição do PB. A continuidade da cadeia da fala é uma de suas características inerentes: desaparecem as fronteiras demarcadoras do final e início das palavras, em virtude das junturas ou sândi externo fechado. Assim, no PB, se a palavra terminar por consoante e a seguinte iniciar por vogal, a consoante migra para a sílaba seguinte. Ex.: /uS/ + /'ɔvuS/ = /'zɔvuS/. Se a palavra terminar pela vogal átona /a/ e a seguinte iniciar por vogal átona, essa assimila /a/ como em /'kaza/ + /IS'kura/ = /kazIS'kura/. O mesmo ocorrerá com sons contíguos que realizam fonemas idênticos, ocasionando a crase como em /'bazI/ + /IS'kura/ = /bazIS'kura/. Todos esses fenômenos afetam os clíticos.

Acresce que as sílabas átonas finais de vocábulo, pela ausência de energia, deixam poucos vestígios para a máquina discriminar de quais sons se trata, crucial no PB, pois o padrão vocabular canônico dos substantivos, verbos e adjetivos são os vocábulos paroxítonos que terminam por átonos /a/, /I/, /U/, seguidos ou não por /S/, ou por /ãw/, /ěj/, /ějS/ (nos três últimos, o centro vocálico nasal contém energia).

É sobre esse *input* que a criança deverá recortar os itens que alimentarão seu léxico inicial. Como o sistema fonológico e a gramática ainda estão em construção, ela não dispõe de nenhuma informação *topdown*, só a que provém da cadeia da fala, para delimitar o item e essa informação é muito opaca.

Por isso, os enunciados iniciais da criança contêm sons que, muitas vezes, distam dos que realizam comumente os fonemas do PB, além de estruturas silábicas inexistentes, como [mbe], fusão da frase 'um beijo', conforme já mencionado.

Existem muitas partículas totalmente ambíguas, que ora funcionam como operadores discursivos, como m, que pode ser um interrogativo, ou negação, ora como muletas de processamento e, nesse caso, descartáveis, sem que a máquina possa decidir.

8 Palavras finais

No decorrer do artigo, examinamos desafios metodológicos à linguística de *corpus* em aquisição da linguagem, muitos dos quais permanecem, apesar das novas tecnologias.

A medida mais importante para definir a fase de aquisição em que se encontra criança, o EME, ainda apresenta problemas, pois muitos critérios foram estabelecidos *ad hoc* para o inglês, carecendo de generalização.

Outro desafio decorre de, na fase 1, início, a gramática da criança apresentar aparência pivotal, com operadores aos quais ela adjunge itens em aberto. Não se podem descartar tais operadores embora pareçam muletas, mero recurso articulatório de desempenho, mas, na verdade, itens gramaticais produtivos.

Segmentar os enunciados da criança em unidades menores persiste como o grande desafio metodológico, mesmo após o uso de novas tecnologias, porque os itens lexicais dela discrepam dos adultos: a criança recorta do *input* sequências que, para ela, ocorrem nos mesmos contextos de uso, com os mesmos significados, baseada no que tem saliência perceptual e sem o conhecimento prévio internalizado do sistema fonológico e gramatical, ainda na fase inicial de construção. Assim, o enunciado [´mbe] não é desmembrado em dois itens, m, que, seria a redução do artigo indefinido, nem bê, a abreviatura de ‘beijo’: ‘bê’ não ocorre isolado, ou combinado com outros itens, com o significado de ‘beijo’, nenhuma vez.

Contudo, acredito ter contribuído para a área quando, em 1974, um ano após a publicação do *A First language*, de Roger Brown, iniciei a coleta de dados de Pé, fiz a transcrição fonética, com os padrões de entoação dos 5530 enunciados da criança, organizei os *corpora*, com os léxicos e os disponibilizei à comunidade científica mundial na Plataforma CHILDES.

Referências

BERNSTEIN RATNER, N.; MACQUINNEY, B. Your Laptop to the Rescue: Using the Child Language Data Exchange System Archive and CLAN Utilities to Improve Child Language Sample Analysis. **Seminars in Speech and Language**, Alphen aan den Rijn, 2016, v. 37, n. 2, p. 74-84. Disponível em: [nan.pdf \(talkbank.org\)](http://nan.pdf(talkbank.org)). Acesso em: 12 abr. 2021.

BLOOM, L. **Language development, form and function in emergent grammars**. Cambridge, Mass.: The M. I. T. Press, 1970.

BLOOM, L. **One word at a time**. The Hague: Mouton, 1973.

BOWERMAN, M. **Early syntactic development**. Londres: Cambridge Univ. Press, 1973.

BROWN, R. A **first language**: The early stages. Cambridge, MA: Harvard University Press, 1973.

CAZDEN, C. B. The acquisition of noun and verb inflection. *In*: FERGUSON, C. A.; SLOBIN, D. J. (Orgs.). **Studies of child language development**. New York: Holt, Rinehart & Winston, 1973, p. 226-240.

CHILDES. **Romance languages**, Scliar-Cabral. Disponível em: <https://childes.talkbank.org/access/Romance>. Acesso em 9 de out. 2020.

CHOMSKY, N. **Aspects of the theory of syntax**. 13. ed. Cambridge, Mass.: The M. I. T. Press, 1964.

FILLMORE, C. J. The case for case. *In*: BACH, E.; HARMS, R. T. (Orgs.). **Universals in linguistic theory**. New York: Holt, Rinehart & Winston, 1968, p. 1-18.

MARTINET, A. **Elementos de linguística geral**. Tradução de Jorge Moraes- Barbosa. Lisboa: Sá da Costa, 1964.

MATLUCK, J. Entonación hispánica. **Anuario de Letras**, v. 5, p. 5–32, 1965.

MENYUK, P. **The acquisition and development of language**. Englewood Cliffs, New Jersey: Prentice-Hall, 1971.

NEWMAN, R. S.; ROWE, M. L.; BERNSTEIN RATNER, N. Input and uptake at 7 months predicts toddler vocabulary: the role of child-directed speech and infant processing skills in language development. **Journal of Child Language**, v. 1, n. 5, ps. 1–16, 2015.

POTTIER, B. **Presentación de la linguística**. Tradução de de Martín Blanco-Alvarez. Madri: Gredos, 1968.

SCLIAR-CABRAL, L. **A explanação linguística em gramáticas emergentes**. 399 f. (Doutorado em Linguística) – Faculdade de Filosofia, Ciências e Letras, Universidade de São Paulo, 1976a.

SCLIAR-CABRAL, L. **A explanação linguística em gramáticas emergentes**. Transcrição fonética (Anexos). 399 f. (Doutorado em Linguística) – Faculdade de Filosofia, Ciências e Letras, Universidade de São Paulo, 1976b.

Recebido em: 21 de outubro de 2020

Aceito em: 5 de abril de 2021

Publicado em maio de 2021

Leonor Scliar-Cabral
E-mail: leonorsc20@gmail.com
ORCID: <https://orcid.org/0000-0003-3163-5482>