CEPLEXicon – um Léxico de Aquisição do Português Europeu[1]

Ana Lúcia Santos, Maria João Freitas*, Aida Cardoso**
* CLUL/FLUL, Universidade de Lisboa

Abstract

CEPLEXicon (version 1.1) is a child lexicon resulting from the automatic tagging of two child corpora: the corpus Santos (Santos, 2006; Santos et al. 2014) and the corpus *Child – Adult Interaction* (Freitas *et al.* 2012), which integrates information from the corpus Freitas (Freitas, 1997). This lexicon represents spontaneous speech produced by seven children (1;02.00 to 3;11.12) during approximately 86h of child-adult interaction. The lexicon contains information pertaining to lemmas and syntactic categories as well as absolute number of occurrences and frequencies in three age intervals: < 2 years; ≥ 2 years and < 3 years; ≥ 3 years.

Keywords/Palavras-chave: acquisition, lexicon, European Portuguese / aquisição, léxico, português europeu

1. Introdução

O CEPLEXicon é um léxico que disponibiliza publicamente dados de produção espontânea de crianças portuguesas. O léxico, na versão 1.1, encontra-se registado na ELRA (ELRA-L0094), correspondendo ao ISLRN 408-817-203-152-3. A produção deste léxico teve como objetivo central fornecer informação sobre o léxico de fala infantil utilizado em situação não experimental de recolha de dados, em situações de interação verbal comuns no quotidiano da criança.

Este léxico infantil foi construído sobre um conjunto de dados cuja recolha resulta de trabalho sobre aquisição do português europeu (PE) como língua materna iniciado, nos anos 90 do século passado, na Faculdade de Letras da Universidade de Lisboa e financiado por diferentes agências públicas de apoio à investigação – para informação relativa ao seu financiamento, consulte-se a página sobre o AcEP (*Acquisition of European Portuguese*), banco de dados que inclui vários recursos para o estudo da aquisição do PE, disponíveis no Centro de Linguística da Universidade de Lisboa (http://www.clul.ul.pt/research-teams/476-acquisition-of-european-portuguese-databank).

Foram usados para construção deste léxico o *corpus* Santos (Santos, 2006; Santos *et al.*, 2014) e o *corpus Child – Adult Interaction* (Freitas *et al.*, 2012), construído a partir do *corpus* Freitas (Freitas, 1997).

A recolha dos dados que estão na base da construção do CEPLEXicon seguiu a tradição dos estudos sobre aquisição de língua materna no campo da linguística. Até ao final dos anos 80, destacavam-se duas linhas de investigação no domínio da aquisição e desenvolvimento linguísticos (Ingram, 1989): uma das linhas, normalmente designada por *Child Language*, incluía trabalhos levados a cabo por psicolinguistas, tendencialmente com base em dados transversais experimentais recolhidos junto de amostras alargadas; a outra linha de investigação, conhecida como *Language Acquisition*, incluía trabalhos produzidos por linguistas, normalmente com base em amostras reduzidas, sendo as crianças avaliadas longitudinalmente em contexto espontâneo. Na primeira perspetiva, o objetivo central era o da disponibilização de informação a partir de produções de um grande número de crianças, com vista à descrição de comportamentos verbais que permitissem identificar os perfis linguísticos de diferentes faixas etárias. Na segunda perspetiva, os objetivos centrais eram os

[1] Este trabalho foi desenvolvido no Centro de Linguística da Universidade de Lisboa (CLUL), no âmbito do projeto *Complement Clauses in the Acquisition of Portuguese* (PTDC/CLE-LIN/120897/2010), financiado pela Fundação para a Ciência e a Tecnologia.

de identificar a arquitetura do conhecimento linguístico e testar análises linguísticas propostas para a gramática-alvo. Os *corpora* subjacentes a Freitas (1997) e a Santos (2006) foram construídos no contexto desta segunda perspetiva de investigação.

Na última década, a associação exclusiva de diferentes metodologias de trabalho a diferentes orientações teóricas tem-se esbatido. Tal veio proporcionar a criação de um novo paradigma de investigação, designado por Ingram (1989) como *Child Language Acquisition*, que permitiu a exportação dos aspetos mais produtivos de cada um dos paradigmas anteriores, no sentido da construção de um conhecimento cada vez mais preciso sobre a aquisição de língua materna. É muito comum, atualmente, a realização de estudos transversais (e até longitudinais) experimentais por parte dos linguistas, com vista ao estudo de uma estrutura específica, que decorre, assim, da adoção de metodologias de recolha típicas dos estudos na área da psicolinguística e permite a testagem de hipóteses linguísticas com base num maior número de sujeitos, em diferentes faixas etárias.

No entanto, os dados longitudinais são, ainda hoje, cruciais para acedermos ao modo como um dado sistema linguístico se vai estruturando gradualmente no cérebro infantil. Desta forma, os *corpora* desta natureza, compilados ao longo de quatro décadas de investigação, muitos deles disponíveis nos sistemas *CHILDES* (MacWhinney, Carnegie Mellon, http://childes.psy.cmu.edu/) e *PHONBANK* (MacWhinney & Rose, Carnegie Mellon & Memorial University of Newfoundland, http://childes.psy.cmu.edu/phon/), como é o caso dos dois *corpora* que foram a base deste trabalho, continuam a ser fontes essenciais de informação para quem pretende avaliar a aquisição gradual de um sistema linguístico, seja com vista à construção de escalas de desenvolvimento de base linguística, seja com o objetivo de estudar o modo como uma dada estrutura linguística emerge e estabiliza no processo de aquisição.

A importância dos dados longitudinais para a construção de conhecimento sobre o desenvolvimento linguístico infantil, particularmente o desenvolvimento lexical, justificou a construção do CEPLEXicon. Uma outra motivação para a construção deste recurso foi a disponibilização em formato acessível de informação sobre o desenvolvimento lexical nos primeiros anos de vida. Muitos dos *corpora* disponíveis publicamente nos dois formatos acima referidos (*CHILDES; PHONBANK*) apresentam ou transcrições fonéticas ou transcrições ortográficas codificadas, que dificultam a procura de informação sobre a natureza e as propriedades do léxico infantil. O que apresentamos é um instrumento que reúne a informação lexical relevante em *corpora* com formatos iniciais distintos e que se pretende de fácil consulta por um leque alargado de profissionais.

2. Construção do CEPLEXicon

2.1. *Corpora* de base

O CEPLEXicon foi desenvolvido com base em dois *corpora* de dados longitudinais espontâneos: o *corpus* Santos (Santos, 2006; Santos *et al*., 2014) e o *corpus Child – Adult Interaction* (Freitas *et al.*, 2012), construído a partir do *corpus* Freitas (Freitas, 1997)[2]. Partindo dos dados destes dois *corpora*, o léxico que aqui se apresenta resulta da anotação morfossintática automática da produção oral de sete crianças com idades compreendidas entre os 1;02.00 e os 3;11.12 anos (Tabela 1), num total de 114 ficheiros de transcrição, cada um correspondendo a sessões de gravação em formato vídeo de 40 a 50 minutos de interação criança-adulto num ambiente naturalista. Os dados tomados como base para a construção deste léxico correspondem assim a cerca de 86h de interação criança-adulto(s).

Crianças	Idade
Inês I.	1;06.06 – 3;11.12
Inês M.	1;05.09 – 2;09.03
Tomás	1;06.18 – 3;10.16
Laura	2;02.30 – 3;03.10
Marta	1;02.00 – 2;02.17
Pedro	2;07.00 – 3;07.24
Raquel	1;10.02 – 2;11.21

Tabela 1: Idades das crianças incluídas nos *corpora* Santos (Santos, 2006; Santos *et al*., 2014) e *Child – Adult Interaction* (Freitas, 1997; Freitas *et al.*, 2012).

[2] Para uma descrição detalhada destes *corpora*, veja-se Santos (2006), Santos *et al*. (2014), Freitas (1997) e Freitas *et al*. (2012).

Refira-se, ainda, que os ficheiros que constituem o *corpus* Santos (Santos, 2006; Santos *et al.*, 2014) foram originalmente transcritos de acordo com o sistema CHILDES (Child Language Data Exchange System) e com recurso ao programa CLAN (MacWhinney, 2000, http://childes.psy.cmu.edu/), estando publicados na base de dados CHILDES (http://childes.talkbank.org/data/Romance/Portuguese/). Já os ficheiros que fazem parte do *corpus Child – Adult Interaction* (Freitas, 1997; Freitas *et al., 2012)* foram inicialmente transcritos foneticamente (Freitas, 1997); com base nestas transcrições fonéticas, foi posteriormente feita uma transcrição em formato *EXMARaLDA* (http://www.exmaralda.org/), seguindo as regras de transcrição ortográfica definidas no Centro de Linguística da Universidade de Lisboa (veja-se a descrição em Freitas *et al.*, 2012).

2.2. Anotação morfossintática prévia dos *corpora*

O léxico CEPLEXicon tem por base a anotação morfossintática dos dois *corpora* anteriormente referidos – o *corpus* Santos (Santos, 2006, Santos *et al.*, 2014) e o *corpus Child – Adult Interaction* (Freitas, 1997, Freitas *et al.* 2012). Esta anotação foi desenvolvida em duas fases: foi realizada num primeiro momento uma anotação automática, à qual se seguiu uma revisão manual (parcial) dos resultados.

Assim, e no que diz respeito à anotação automática, esta foi realizada com recurso ao etiquetador desenvolvido no Centro de Linguística da Universidade de Lisboa (Généreux, Hendrickx & Mendes, 2012), sendo de notar que este etiquetador foi inicialmente treinado estatisticamente com dados de *corpora* de português escrito. Assim sendo, considerou-se necessário adaptá-lo a dados de produção oral e, em particular, à produção de crianças, através de regras específicas, tal como descrito em Santos *et al.* (2014). A relevância deste trabalho prévio de adaptação do etiquetador automático encontra-se refletida na avaliação realizada sobre os resultados da sua aplicação ao *corpus* Santos, já que esta avaliação revelou uma taxa de precisão de 94,9% para a etiquetagem morfossintática e de 98% para o lematizador (Santos *et al.*, 2014), resultados que se encontram dentro dos parâmetros esperados.

O conjunto de etiquetas POS (*part-of-speech*) usado na anotação automática foi o utilizado na anotação de outros *corpora*, nomeadamente a anotação do CRPC[3] (Généreux, Hendrickx & Mendes, 2012), o que garante um elevado grau de uniformidade na anotação de *corpora* e, além disso, assegura a adequação do conjunto de etiquetas agora usado.

Como resultado desta tarefa de anotação morfossintática, cada ficheiro de transcrição tem uma fiada de classificação morfossintática gerada automaticamente pelo etiquetador. Nesta fiada, encontram-se o lema e a etiqueta POS atribuídos a cada uma das palavras presentes na fiada da transcrição, como exemplificado em (1).

(1) *MAE: e mais?
%xmor: CJ|e ADV|mais ?
*CHI: e pa(ra) a p(r)aia.
%xmor: CJ|e PREP|para DA|a CN|praia .

[Tomás 2;4.0]

Como se pode observar em (1), é atribuída a cada palavra uma etiqueta POS correspondente a uma categoria morfossintática (*e.g.*, "CN" é a etiqueta usada para codificar nomes comuns – C(ommon) N(oun)), seguida de um lema (*e.g.*, *praia*). A lista completa das etiquetas POS usadas na anotação morfossintática dos *corpora* e que foram mantidas no léxico pode ser consultada no Anexo 1 (mais informação em 2.3). Importa ainda referir que, tal como descrito em Santos *et al.* (2014), algumas anotações específicas e metadados incluídos na transcrição ortográfica dos ficheiros foram removidos ou ignorados no processo de geração da fiada morfossintática (*e.g.*, símbolos usados para denotar discurso ininteligível, como "xxx"), como mostra o exemplo (2).

(2) CHILD: xxx que(r) bo(n)eca.
%xmor: V|querer CN|boneco .

[Inês 2;3.22]

Posteriormente, e com o objetivo de preparar uma versão final do léxico, foi levada a cabo uma revisão manual parcial das 98200 palavras anotadas pelo etiquetador e que incidiu sobre a informação respeitante ao lema e respetiva etiqueta POS. Ressalve-se, contudo, que se tratou de uma revisão parcial dos dados, pelo que se descrevem, em seguida, os principais casos em que se considerou necessário proceder a uma verificação contextual e, em alguns casos, correção da etiquetagem morfossintática nos ficheiros de transcrição.

[3] http://www.clul.ul.pt/en/resources/183-crpc.

Em primeiro lugar, foram corrigidos os casos em que foi possível identificar inequivocamente um erro do etiquetador, quer a nível do lema quer a nível da etiqueta POS. Assim sendo, por exemplo ocorrências como "V|aleija" foram alteradas para "V|aleijar", uma vez que o lema da categoria Verbo consiste na forma infinitiva, e classificações como "ADJ|banheira" foram corrigidas para "CN|banheira", pois "banheira" é um nome comum e não um adjetivo.

Em segundo lugar, alguns casos conhecidos e frequentes de ambiguidade (verbo / nome; nome / adjectivo) foram verificados nos ficheiros de transcrição e a etiqueta POS foi determinada com base no contexto. São exemplo deste tipo de verificação ocorrências como "CN|colar" em relação às quais é necessário verificar se a anotação está correta, tratando-se efetivamente de um nome comum, ou se, pelo contrário, a anotação terá de ser alterada para "V|colar", por se tratar de um verbo. No entanto, devido a restrições a tempo, palavras com uma elevada frequência no *corpus*, como *a*, que pode corresponder a uma preposição, ao determinante artigo definido feminino singular ou ao pronome pessoal clítico acusativo feminino singular, não tiveram a sua etiquetagem verificada. Neste caso, assumiu-se que a elevada taxa de precisão atingida pelo etiquetador neste *corpus* seria suficiente para nos fazer assumir uma elevada taxa de precisão também nestes casos. Qualquer trabalho que se centre especificamente nestas categorias pode complementarmente usar como fonte direta os *corpora* que aqui foram tomados como base.

Finalmente, as palavras associadas a etiquetas POS que se podem considerar pouco comuns na produção de crianças com idade inferior a quatro anos foram também verificadas nos ficheiros de transcrição e corrigidas sempre que necessário. Neste sentido, e por se considerar, por exemplo, que seria mais provável que uma criança produzisse palavras como "qual" e "onde" como interrogativos e não como relativos, todas as ocorrências de "REL|qual" e de de "REL|onde" foram verificadas contextualmente.

Não obstante, deve-se ter em conta que há uma taxa de erro associada a qualquer tarefa de anotação e uma anotação automática não é exceção. Isto significa que qualquer utilização deste léxico terá de ter em conta que se trata de um léxico construído com base num *corpus* automaticamente anotado e que foi objeto apenas de uma revisão manual parcial. Importa, ainda assim, relembrar que a taxa de acerto na aplicação deste etiquetador automático ao *corpus* de Santos (2006) se encontra dentro dos parâmetros esperados (Santos *et al*., 2014).

2.3. Estrutura do CEPLEXicon

O CEPLEXicon resulta da anotação automática de 98200 palavras, tal como descrito anteriormente. Desta anotação foi extraído um total de 2201 lemas, incluindo 1043 nomes comuns, 130 adjetivos e 303 verbos.

A concatenação dos diferentes lemas em diferentes categorias sintáticas, partindo das etiquetas que constituem o *output* do etiquetador e descritas em Généreux, Hendrickx & Mendes (2012), foi ainda objeto de algumas decisões, de que destacamos as seguintes (para mais informações, consulte-se o manual associado ao CEPLEXicon):

(i) o etiquetador atribui diferentes etiquetas POS a diferentes formas verbais (por exemplo, distingue infinitivos, gerúndios e particípios passados). Todas estas categorias foram reagrupadas como V(erbo) no léxico aqui apresentado;

(ii) sequências como "Aquário Vasco da Gama" foram considerados como uma única unidade, um nome próprio.

Assim, a lista de etiquetas POS que se encontra no anexo 1 diz respeito apenas à lista das que foram mantidas neste léxico.

No que diz respeito ao formato em que se encontra disponível, o CEPLEXicon pode ser consultado em formato .xls, permitindo usar todas as funcionalidades deste tipo de ficheiro (incluindo filtros). O léxico contém as seguintes informações distribuídas por diferentes campos (colunas), como se descreve em seguida.

1) Lista de palavras (lemas) produzidas pelas sete crianças, ordenada alfabeticamente.

2) Categoria POS correspondente a cada lema.

3) Número absoluto de ocorrências de cada lema por intervalo de idade: <2 anos; ≥ 2 anos e < 3 anos; ≥ 3 anos.

4) Frequência em percentagem de cada lema por intervalo de idade: <2 anos; ≥ 2 anos e < 3 anos; ≥ 3 anos.

5) Idade da primeira ocorrência de cada lema em cada criança (ano, mês e dia).

6) Observações.

O campo denominado "Observações" inclui informações sobre a classificação de determinadas palavras como formas familiares ou estrangeirismos.

3. Considerações finais

O CEPLEXicon é um instrumento posto à disposição da comunidade que permite recolher informação sobre o léxico (incluindo lemas e categorias sintáticas) no desenvolvimento linguístico monolingue entre um ano e os quatro anos de idade. A sua utilização pode ser relevante em várias áreas, incluindo, pelo menos, as seguintes:

(i) construção de materiais de avaliação e intervenção em contexto clínico;

(ii) preparação de materiais didáticos, para uso em ambiente escolar;

(iii) criação de materiais de natureza lúdica ou literária (jogos, livros infantis, aplicações informáticas ou outros).

Neste momento, é já utilizado como referência na construção das versões portuguesas dos questionários desenvolvidos no âmbito do projeto *Tracking Studies and Validation of the MacArthur-Bates Communicative Development Inventories for European Portuguese* (PTDC/MHC-PED/4725/2012, FCT, COMPETE e FEDER), sediado na Universidade do Minho.

A utilização do CEPLEXicon é gratuita, sendo, contudo, mediada pela ELRA (http://catalog.elra.info/product_info.php?products_id=1244) e obrigando à citação explícita de fontes, de acordo com o estabelecido no manual e no contrato com esta instituição.

Referências

Généreux, M., I. Hendrickx & A. Mendes (2012). Introducing the Reference Corpus of Contemporary Portuguese On-Line. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation - LREC 2012*. European Language Resources Association (ELRA), pp. 2237-2244.

Freitas, M. J. (1997). *Aquisição da estrutura silábica do Português Europeu*. Ph.D. Dissertation. Universidade de Lisboa.

Freitas, M. J., A. Tanganho, M. Rocha & P. Oliveira (2012) Child-Adult Interaction: A Database on European Portuguese, CLUL, Anagrama.

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

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* Mahwah / New Jersey: Lawrence Erlbaum Associates, 3rd Edition.

Santos, A. L. (2006). *Minimal Answers. Ellipsis, Syntax and Discourse in the Acquisition of European Portuguese*. Dissertação de Doutoramento. Universidade de Lisboa. (Publicado em 2009, Amsterdam / Philadelphia: John Benjamins).

Santos, A. L., M. Génereux, A. Cardoso, C. Agostinho, S. Abalada (2014) A corpus of European Portuguese child and child-directed speech. In *Proceedings of the 9th Conference on Language Resources and Evaluation – LREC 2014*. European Language Resources Association (ELRA).

Anexo 1

Lista de etiquetas POS incluídas no CEPLEXicon

Etiqueta	Categoria	Exemplos
ADJ	Adjectivos	bom, brilhante, eficaz, …
ADV	Advérbios	hoje, já, sim, felizmente, algo, …
CARD	Cardinais	zero, dez, cem, mil, …
CJ	Conjunções	e, ou, mas, porque, pois…
CL	Clíticos	o, lhe, se, …
CN	Nomes Comuns	computador, cidade, ideia, …
DA	Artigos Definidos	o, os, a, as.
DEM	Demonstrativos	este, esses, aquele, …
DFR	Denominadores de Fracções	meio, terço, décimo, %, …
DM	Marcadores Discursivos	pronto, enfim, pá, …
EXC	Exclamativos	que, quanto, ...
IA	Artigos Indefinidos	uns, umas, …
IND	Indefinidos	tudo, alguém, ninguém, …
INT	Interrogativos	quem, como, quando, …
ITJ	Interjeições	bolas, caramba, olá, fogo, alto, …
LTR	Letras	a, b, c, …
MGT	Classes de Magnitude	unidade, dezena, dúzia, resma, …
MTH	Meses	Janeiro, Dezembro, …
ORD	Ordinais	primeiro, centésimo, penúltimo, …
PADR	Partes de Endereços	Rua, av., rot., …
PNM	Partes de Nomes	Lisboa, António, João, …
POSS	Possessivos	meu, teu, seu, …

PPA	Particípios Passados (não em tempos compostos)	pintado, afirmados, vivida, …
PREP	Preposições	de, para, desde, em, …
PRS	Pronomes Pessoais	eu, tu, ele, …
QNT	Quantificadores	todos, muitos, nenhum, …
REL	Pronomes Relativos	que, cujo, quem, …
STT	Títulos	Presidente, dr., prof., …
UM	"um" ou "uma"	um, uma
UNIT	Unidades de Medida (formas abreviadas)	Kg, h, seg, Hz, Mbytes,...
VAUX	Verbos "ter" ou "haver" em tempos compostos	temos, haveriam, …
V	Verbos	falou, falaria, …
WD	Dias da Semana	segunda, terça-feira, sábado, …