UMA FERRAMENTA PARA A PESQUISA EM CORPORA DE AQUISIÇÃO DE LINGUAGEM

1. INTRODUÇÃO

Corpora de linguagem dirigida a e produzida por crianças são recursos valiosos para estudos de aquisição da linguagem, provendo a base para o desenvolvimento de hipóteses sobre a aquisição, bem como comparações e avaliações. Para estudos computacionais esses corpora podem ser utilizados como uma aproximação do ambiente linguístico ao qual uma criança é exposta [5]. Neste contexto, o CHILDES [6] é uma base de corpora amplamente utilizada por apresentar a transcrição da fala de crianças e por possuir dados em mais de vinte e cinco línguas. Os dados utilizados neste trabalho são dos corpora em português disponíveis no CHILDES: Batoréo [1] (com narrativas de adultos e 30 de crianças), Porto Alegre [4] (dados longitudinais e transversais de crianças entre 5 e 9 anos) e Florianópolis (estudo longitudinal de uma criança).

Estudos focados nos dados em inglês motram que crianças tendem a utilizar esquemas canônicos de sentenças [9], assim como, omitir partes iniciais destas [10] e, ainda, que a frequência desempenha um papel importante no processo de aquisição [11].

Além dos corpora são importantes as ferramentas que possibilitam o manuseio destes de forma a facilitar a interação e extração de informações. Esse trabalho apresenta o sistema Child Language Database ([12] e [13]) para o Português que provê anotações sintáticas, semânticas e psicolinguísticas para as sentenças originais do CHILDES e possibilita buscas complexas envolvendo combinações dos diferentes níveis de anotação. O desenvolvimento deste tipo de sistema, para a língua inglesa, é descrito em [2]. Estendendo o trabalho de Wilkens e Villavicencio ([12] e [13]) esse trabalho apresenta o processo de anotação e organização da dados para o português, assim como, a incorporação destes dados na interface Child Language Database. Com isto pretendemos prover um sistema que permite a identificação de padrões, tanto para a analise de sentenças quanto para a geração de gráficos sumarizando as distribuições dos padrões, permitindo a avaliação de hipóteses sobre a aquisição de linguagem de forma rápida e fácil. Na seção 2 descrevemos os recursos e métodos utilizados no processo de anotação, e a ligação dos dados anotados com a interface de consulta. Na seção 3 descrevemos o resultado do processo de anotação e apresentamos alguns dos padrões presentes no corpus.

2. MATERIAIS E MÉTODOS

A fim de propiciar um recurso com busca integrada realizamos o processo de anotação, assim como, a incorporação dos dados anotados à interface Child Language Database. No processo de anotação utilizamos um analisador sintático e recursos eletrônicos contendo informações psicolinguísticas e semânticas. O analisador sintático utilizado foi o PALAVRAS [2] que por ser robusto sempre retorna uma análise, mesmo que a frase seja incompleta ou não gramatical. Este possui uma acurácia de 99% para a classificação de classes morfossintáticas, de 96% para analise sintática e de 91,8% para termos compostos [2]. Além das informações sintáticas as informações psicolinguísticas são relevantes no estudo da aquisição da linguagem. Utilizamos os dados apresentados em Cameirão e Vicente [3], que coletaram informações psicolinguísticas para 1749 palavras: da idade de aquisição,

comprimento da palavra, densidade da vizinhança, frequência, familiaridade, imaginabilidade e concretude), usando 685 pessoas como fonte dos dados. Para os dados semânticos utilizamos a Wordnet.PT [14], que é uma base do Português estruturada como uma rede léxico-conceptual em torno de um conjunto de relações como sinonímia e hiponímia. A versão atual contém cerca de 19.000 expressões, repartidas por vários campos semânticos.

O processo de anotação consistiu de quatro passos: (1) remoção das meta-anotações originais do CHILDES, tais como palavras incompletas, anotações fonológicas e notas de transcrição; (2) anotação das classes sintáticas das palavras e das árvores de dependência das sentenças, com o analisador PALAVRAS; (3) anotação das informações psicolinguísticas das palavras; e (4) anotação das relações semânticas da Wordnet.PT, incluindo as palavras relacionadas, e a quantidade delas.

Para a inclusão dos dados anotados no Child Language Database foi criado um banco de dados. O primeiro passo para a inserção no banco a criação de uma tabela com as seguintes informações para cada sentença: nome original do corpus (Batoréo, Porto Alegre ou Florianópolis), dados do falante (nome, idade, gênero, grupo, papel, grau de educação) e frases anotadas (frase original, frase com anotação morfológica, frase com analise de dependência e frase com outras anotações que o analisador disponibiliza). Após foi criada outra tabela com anotações em nível de palavra, identificada na sua forma canônica e anotada com classe gramatical, dados psicolinguísticos e semânticos, e frequências de ocorrências por idade da forma superficial, forma canônica e classe gramatical. Desta forma os dados anotados estão no formato utilizado pela interface web, assim permitindo a qualquer usuário escolher combinações de campos do banco de dados para realizar buscas.

3. RESULTADOS

Nesta seção descrevemos os resultados obtidos pelo processo de anotação dos dados e exemplificamos algumas das buscas e análises que podem ser realizadas através da Child Language Database.

Após o processo de anotação obtivemos no total 32924 palavras das quais 8725 palavras distintas, 23078 sentenças e 64 falantes. Todas as palavras estão anotadas com as classes gramaticais, onde 23.211 são substantivos, 2.246 nomes próprios, 31.268 verbos, 2.647 adjetivo e as 96018 restantes formadas por outras classes. Tabela 1 lista a frequência de cada classe gramatical por idade.

Tabela 1: Frequência no corpus das classes gramaticais por idade.

Classe Gramatical	1 ano	2 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos
Substantivos	3968	1445	117	3144	4173	4830	5241	293
Nome próprio	316	94	18	393	414	529	453	29
Verbo	1856	2170	228	5152	6355	7566	7601	340
Adjetivo	315	124	7	421	559	550	636	35
Advérbio	1458	748	163	2746	3869	4346	4696	178
Determinante	1035	974	175	4499	5671	6953	6777	294
Pronome independente	74	196	19	405	644	711	663	17
Preposição	326	308	143	2913	4022	5452	5351	245
Conjunção subordinativa	12	54	38	428	525	593	710	32

Número	231	78	42	432	508	520	653	39
Conjunção coordenativa	10	182	54	1525	1820	2087	1958	102
Pronome pessoal	440	336	114	2504	3394	3888	3981	162
Interjeição	656	473	9	303	289	328	396	27
Elemento composto	0	0	0	3	1	2	8	0
Total	2312	1349	61	1275	1176	1035	1261	92

Das 1749 palavras contempladas por Cameirão e Vicente [3] apenas 1400 estão presentes no corpus e foram anotadas com esses dados, sendo 828 verbos e 386 substantivos distintos e as restantes 186 são de outras classes. Essas tem uma a frequência de ocorrência no corpus de 3079 sentenças para os verbos e 1570 para os substantivos e 899 para as restantes, apresentando uma idade de aquisição média de 3,4 (min=1,23, max=7,83), familiaridade de 1,78 (min=1,11, max=3,33), concretude de 4,89 (min=2, max=6,8) e imaginabilidade 4,8 (min=2,23, max=6,77).

Os dados da Wordnet.PT [14] possibilitaram a anotação de 8409 palavras sendo 2416 palavras distintas divididas em 4134 substantivos (1065 substantivos distintos), 3560 verbos (1134 verbos distintos), 693 adjetivos (210 adjetivos distintos) e 22 advérbios (7 advérbios distintos).

Com base nesses dados foram analisadas as evoluções de três variáveis nos corpora: classes gramaticais, idade de aquisição e grau de polissemia.

Quanto ao uso das classes gramaticas durante o desenvolvimento da criança, pela analise dos dados, identifica-se que inicialmente (com menos de dois anos) as crianças tendem a utilizar predominantemente mais substantivos (quase metade de seu vocabulário), o que é compatível com o 'bias do substantivo' proposto por Tardiff [15], entre outros. Com o aumento da idade o uso das outras classes gramaticais tende a aumentar, e, por exemplo, aos dois anos a frequência dos verbos ultrapassa a dos substantivos. Além disso, a classe fechada de palavras (determinantes, preposições e pronomes) apresenta uma frequência crescente até os três ou quatro anos e após a frequência estabiliza.

Em termos da idade de aquisição, percebe-se em todas as idades um uso predominante das palavras avaliadas como adquiridas aos dois anos, e apesar do que é sugerido pelo índice, o seu uso proporcional maior, entre todas as palavras, é nas crianças com dois anos ou menos. Essa divergência se repete nas palavras com índice de idade de aquisição de quatro anos ou menos que apresentam um maior uso proporcional nas crianças com dois anos ou menos, sendo praticamente constante nas demais. Ressaltamos que estes resultados podem sofrer influência do número reduzido de palavras no corpus para o qual se tem informação do índice de idade de aquisição.

Para a análise da polissemia utilizamos como indicativo o número de hiperônimos da palavra, onde é identificado um aumento no uso de palavras com mais hiperônimos conforme aumenta a idade. Ou seja, com o passar da idade das crianças os dados indicam um uso de palavras mais polissemicas. Contudo, destacamos que em todas as idade há um baixo uso de polissemia (menor que 5 hiperônimos) e com mais de cinco anos aumenta o uso de palavras com alta polissemia (mais de 20 hiperônimos).

4. CONCLUSÕES E TRABALHOS FUTUROS

Nesse artigo descrevemos o processo de anotação de corpora com transcrições de falas de crianças, com informações sintáticas, semânticas e psicolinguísticas. Os corpora anotados foram disponibilizados através de uma interface Web para consulta e geração textual e gráfica dos dados. Além disso, foi feita uma descrição dos corpora e uma análise dos dados de três variáveis de acordo com os diferentes grupos etários. Como trabalhos futuros se pretende analisar os comportamentos dos dados ao longo dos anos e possibilitar buscas predefinidas na interface Web. Prevê-se também a integração com os dados dos corpora em inglês do CHILDES apresentado em [15].

1. Batoréo,H. 2000. Expressão do Espaço no Português Europeu. Contributo psicolinguístico para o Estudo da Linguagem e Cognição. Dissertação de doutorado, Fundação Calouste Gulbenkian e Fundação para a Ciência e a Tecnologia, Ministério da Ciência e da Tecnologia, Lisboa
2. Bick, E. 2000. The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. [S.l.]: University of Arhus.
3. Cameirão, M.L. and Vicente, S.G. 2010. Age-of-acquisition norms for a set of 1,749 portuguese words. Behavior research methods 42, Springer.
4. Guimarães, A. M. 1994. Desenvolvimento da linguagem da criança na fase deletramento. Cadernos de Estudos Linguísticos, 26, 103-110
5. Wintner, S. 2010. Computational Models of Language Acquisition. CICLing'10.
6. MacWhinney, B. (1995). The CHILDES project: Tools for analyzing talk (2nd ed.). Lawrence Erlbaum Associates.
7. Goldberg, Adele E. (1999). The Emergence of Language, chapter Emergence of the semantics of argument structure constructions, pages 197–212. Carnegie Mellon Symposia on Cognition Series.
8. Valian, V. 1991. Syntactic subjects in the early speech of American and Italian Children. Journal of Cognition.
9. Slobin, D., Bever, T. (1982) Children use canonical sentence schemas. Cognition, 12:229–265.
10. Freudenthal, D., Pine, J., Gobet, F. (2006) Modelling the development of children′s use of optional infinitives in Dutch and English using MOSAIC. Cognitive Science, 30:277–310.
11. Bybee, J. Regular morphology and the lexicon. Language and Cognitive Processes, 10(5):425–455, 1995.
12. Wilkens, R., Proença, M., Villavicencio, A., (2012). An Environment for searching Portuguese child language corpora. International Conference on Computational Processing of the Portuguese Language (PROPOR)
13. Wilkens, R. (2012). Searching the Annotated Portuguese Childes Corpora. In Proceedings of Conference of the European Chapter of the Association for computational Linguistics (EACL) Workshop of cognitive aspects of computational language acquisition and loss.
14. Marrafa, P. (2002) Portuguese Wordnet: general architecture and internal semantic relations. DELTA: Documentação de Estudos em Linguíistica Téorica e Aplicada 18 (SPE), 131–146
15. Tardif, T., Gelman, S., Xu, F. (1999) Putting the "noun bias" in context: a comparison of English and Mandarin. Child Development 70:620-635.

16. Villavicencio, A., Yankama, B., Berwick, R., Idiart, M. (2012) A large scale annotated child language construction database. In Proceedings of the 8th LREC, Istanbul, Turkey.