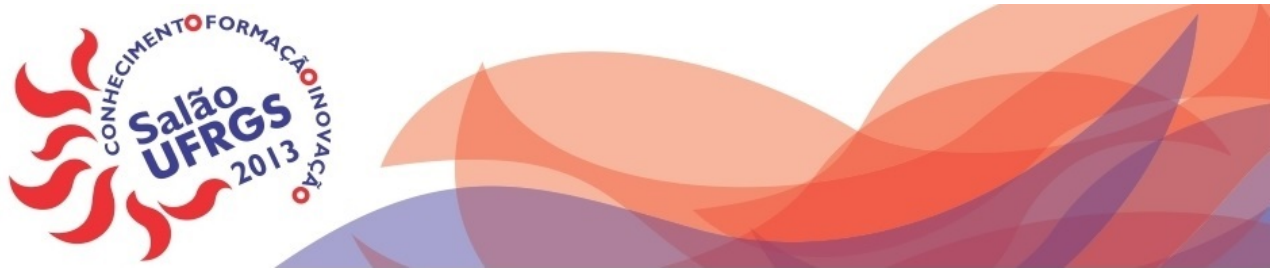

Evento	Salão UFRGS 2013: SIC - XXV SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2013
Local	Porto Alegre - RS
Título	Construção de um banco de dados para análise comparativa de linguagem
Autor	CLEI ANTONIO DE SOUZA JUNIOR
Orientador	MARCO AURELIO PIRES IDIART

O estudo da linguagem nos auxilia a entender a cognição humana, nesse estudo uma área desafiadora para diversas disciplinas é a aquisição de linguagem, isto em computação é estudado através do processamento de linguagem natural. Para tanto uma tarefa importante é a identificação das variáveis do processo de aquisição (neste trabalho é abordado apenas a aquisição lexical). Para tal, foram escolhidas métricas psicolinguísticas para associar a um corpus (conjunto de textos) e foram realizadas análises estatísticas sobre esses dados.

Utilizamos as palavras do corpus CHILDES [MacWhinney, 2000], que contém transcrições de diálogos de crianças de diversas idades Destes dados utilizamos as idades menores que 4 anos (devido a quantia de dados) e corpora longitudinais (os dados de uma criança foram comparados com dados dela mesma quando maior). Para este estudo, foram utilizados os dados da língua inglesa devido a maior quantia de fontes para anotação deles.

As palavras foram anotadas com informações sintáticas (vindas da Valex [Korhonen et al. 2006], FrameNet [Baker et al 1998] e VerbNet [Kipper et al. 2008]), semânticas (da WordNet [Miller 1995]), e psicolinguísticas (do MRC [Wilson 1988]), utilizando apenas 4 classes gramaticais (verbo, substantivo, adjetivo e advérbio).

A fim de identificar quais informações anotadas são diferentes entre as idades utilizamos a Correlação de Spearman para identificar os dados com baixa correlação entre pares de idade. Consideramos as métricas que apresentaram correlação entre -0.3 e 0.3, são valores nesse intervalo que indicam baixa correlação.

Foram obtidos resultados que mostram diferença entre as idades para polissemia dos verbos para diversos pares de idades e em para as métricas de antônimos dos adjetivos (antônimos diretos e indiretos) para um par de idades (1 e 2 anos).

Através da criação desse banco de dados com as palavras anotadas, houve uma contribuição para a formação de uma base empírica de dados que podem ser utilizada para outros tipos de análise de palavras. Das análises realizadas foi possível entender melhor o aprendizado da linguagem e trabalhos de classificação de crianças em idades tendo como base apenas as palavras utilizadas por elas podem ser realizados.

Referências:

MACWHINNEY, B. The CHILDES Project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

MILLER, George A. WordNet: A Lexical Database for English. Communications of the ACM. Vol. 38, No. 11: 39-41, 1995.

FELLBAUM, Christiane. Wordnet: An Eletronic Lexical Database. Cambridge, MA: MIT Press, 1998.

BAKER, Collin F., FILMORE, Charles J., LOWE, John B. The Berkeley FrameNet Project. In Proceedings of COLING-ACL'98, pages 86–90, Montréal,Canada, 1998.

KIPPER, Karin; KORHONEN, Anna; RYANT, Neville; PALMER, Martha. A Large-scale Classification of English Verbs. In Language Resources and Evaluation Journal Vol. 42(1), pp. 21-40, Springer Netherland, 2008.

WILSON, M.D. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. Behavioural Research Methods, Instruments and Computers, 20(1), 6-11, 1988

KORHONEN, Anna; KRYMOLOWSKI; BRISCOE, Ted. VALEX Disponível em: http://www.cl.cam.ac.uk/users/alk23/subcat/lexicon.html .