Processamento de Linguagem Natural:

Análise de Expressões Multipalavra

Maitê Friedrich Dupont
mfdupont@inf.ufrgs.br

Profª Aline Villavicencio
avillavicencio@inf.ufrgs.br

Carlos Ramisch
ceramisch@inf.ufrgs.br

O que é?

Expressões multipalavra (EMs): combinações de palavras que apresentam idiossincrasias lingüísticas ou estatísticas

* Verbos frasais: carry up, consist of
* Verbos de suporte: tomar um banho
* Compostos: carro de polícia, bode expiatório
* Expressões idiomáticas: engolir o sapo, dar para trás

mwetoolkit (mwetoolkit.sf.net):
ferramenta automatizada para a identificação e extração de EMs a partir de corpora utilizando métodos estatísticos.

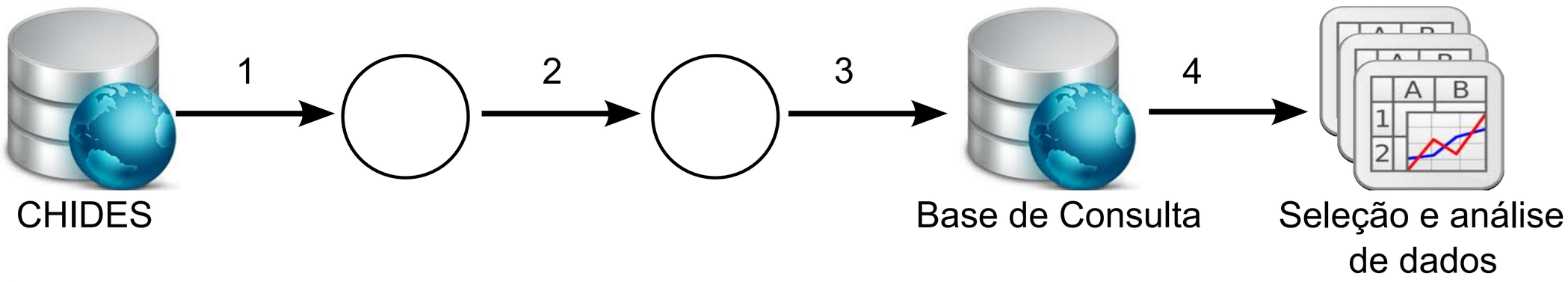
Motivação

A modelagem computacional das linguagens humanas tem o potencial de avançar o entendimento sobre os processos cognitivos relacionados a linguagem. Permite investigar alternativas para aquisição, organização, processamento e dissolução da linguagem, e examinar fatores que podem influir em cada um destes processos, por exemplo, determinando a ordem com que palavras são incorporadas no vocabulário.

Objetivo

Investigar o desenvolvimento da linguagem com foco em Expressões multipalavra (Ems) do tipo VPC (verpo-partícula) buscando formas de categorizar as diferentes etapas do processo de obtenção da língua.

Trabalho Realizado



1. Anotação e parse:
O corpus foi anotado pelo parser RASP de forma a identificar os papéis gramaticais

2. Extração de EMs
O corpus foi processado pela ferramenta mwetoolkit de forma a obter um levantamento dos VPCs com até 5 palavras de distância presentes no corpus.

3. Inserção em um banco de dados
Os dados foram inseridos em um banco de dados de forma a facilitar as consultas e obtenções de datos

4. Separação e triagem
Os dados foram separados de acordo com a faixa etária do falante em 4 grupos: 0-24 meses, 25-48 meses, 49-72 meses e 73-96 meses.

CHILDES	
Total de frases anotadas:	482 137
% de frases com VPC:	9,19%
% de frases com VPC após limpeza dos dados:	7,95%
% de frases com VPC de falantes entre 0 e 24 meses:	7,30%
% de frases com VPC de falantes entre 25 e 48 meses:	68,24%
% de frases com VPC de falantes entre 49 e 72 meses:	20,97%
% de frases com VPC de falantes entre 73 e 96 meses:	3,49%

Conclusões

* Os dados utilizados no experimento não são 100% confiáveis, pois são transcritos automaticamente. O processo de limpeza ajuda a eliminar diversas inconsistências.
* Existe carência de dados de algumas faixas etárias em relação a outras, e portanto foi feita também uma amostragem dos valores.
* As 10 partículas mais utilizadas em VPCs por crianças de qualquer faixa etária são as mesmas utilizadas por adultos.
* Os 10 verbos mais utilizados em VPCs por crianças de qualquer faixa etária são praticamente os mesmos.
* Os 10 VPCs mais utilizados pelas diferentes faixas etárias são semelhantes aos dos adultos, mas apresentam diferenças significativas