A modelagem computacional das linguagens humanas tem o potencial de avançar o entendimento sobre os processos cognitivos relacionados a linguagem. Em particular permite investigar possíveis 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.

Dentro desse contexto, o objetivo desse trabalho é investigar o desenvolvimento da linguagem com foco em Expressões multipalavra (Ems) buscando formas de categorizar as diferentes etapas do processo de obtenção da língua. EMs são combinações de palavras que apresentam idiossincrasias morfológicas, sintáticas, semânticas, pragmáticas ou estatísticas. EMs são uma característica importante das línguas humanas, e incluem fenômenos tais como verbos frasais (*carry up, consist of*), verbos de suporte (*tomar um banho, dar uma caminhada*), compostos (*carro de polícia, bode expiatório*) e expressões idiomáticas.

O presente trabalho consiste na análise de frases de crianças retiradas de uma base contendo interações naturais transcritas entre adultos e crianças, a CHILDES, a fim de determinar semelhanças e diferenças entre o vocabulário de Ems de diferentes grupos etários. CHILDES foi estabelecida em 1984 para servir de repositório para dados de aprendizagem de primeira língua, possuindo, hoje, dados de mais de 20 idiomas distintos. Dividimos as frases para análise em grupos de frases de crianças de 0 a 24 meses, 25 a 48 meses, 49 a 72 meses e 73 a 96 meses. Para efetuar a análise, utilizamos uma ferramenta automatizada para identificação e extração de EMs a partir de corpora, o mwetoolkit [Ramisch et al. (2010a)].

Este tipo de análise pode servir de base para que possamos, por exemplo, comparar padrões de uso esperados de uma EM para uma faixa etária, com o seu uso efetivo por um falante, possibilitando a detecção de possíveis desvios no aprendizado. Também poderemos melhorar os algoritmos de tradução automática ao definir a idade do autor e que palavras da segunda língua são mais apropriadas e condizentes com a faixa etária, e também auxiliar, por exemplo, na educação e aprendizado de linguagem, indicando ao usuário os textos mais adequados à sua "idade vocabular".

Concomitantemente, buscamos expandir as funcionalidades da ferramenta mwetoolkit, de forma que a mesma possa ser utilizada em uma maior diversidade de estudos, provendo mais informações, ao incluir na sua distribuição mais um algoritmo auxiliar que converte a saída de um parser para o formato de entrada do mwetoolkit.