A manipulação correta de expressões compostas (ECs) é um dos grandes problemas enfrentados no processamento de linguagens naturais, e é importante por exemplo, em sistemas de tradução automática de textos. A extração automática de ECs e inserção em dicionários é de suma importância já que dicionários existentes contem apenas uma pequena parte dessas expressões. Esta pesquisa trata da aquisição automática de um tipo particular de ECs, as construções verbo-e-partícula (VPCs) – como, por exemplo, "take off" em "Our plane took off late" – na língua inglesa, a partir de textos escritos, com base nas suas propriedades estatísticas e regularidades lingüísticas.

A técnica proposta para essa tarefa combina informações estatísticas calculadas a partir da ocorrência de VPCs em textos, com os padrões sintáticos em que elas ocorrem e informações semânticas sobre os seus sinônimos. A construção automática de um dicionário de sinônimos a partir de uma coleção de textos é feita através da determinação de agrupamentos de palavras com usos semelhantes, através de medidas estatísticas sobre as relações gramáticas entre as palavras extraídas desses textos. Além disto, são usadas informações multilíngües sobre VPCs (e construções equivalentes) extraídas a partir de textos equivalentes em várias línguas.

Neste estudo estão sendo usados os textos em Português e Inglês do Europarl, um conjunto de textos formado por mais de 30 milhões de palavras com textos do Parlamento Europeu. A partir da parte inglesa do Europarl, o experimento resultou em dois dicionários de sinônimos com mais de 5000 verbos (cada um relacionado com em média 4000 outras palavras). Esse sistema e esses dados serão usados em experimentos para a determinação de VPCs, e de suas características sintáticas e semânticas.