Abordagens recentes do campo da linguística computacional se propõem a representar relações semânticas entre palavras a partir de relações lineares entre vetores multidimensionais representando aquelas palavras. Esse tipo de representação remonta a abordagens da chamada semântica distribucional, perspectiva de estudo do significado a partir da distribuição das palavras em coleções de dados linguísticos, cuja máxima básica é: palavras com significados parecidos ocorrem em contextos parecidos.

A explosão de dados e o aumento da capacidade de processamento disponíveis tornaram possível a implementação desse tipo de modelo de modo que seu funcionamento global não é interpretável para seres humanos, ao mesmo tempo em que obtém resultados interessantes em tarefas distintas de processamento de linguagem natural.

Essa comunicação tem a intenção de apresentar os fundamentos desse tipo de representação do significado em sua roupagem mais recente: métodos de contagem, matrizes documento por termo e métodos de cálculo de distância entre vetores. Também serão elencados alguns questionamentos pontuais e fundamentais que a linguística pode trazer para esse tipo de abordagem.

A partir desse ponto, apresentarei alguns exemplos das aplicações das medidas de similaridade entre os vetores e como esses resultados tornam possíveis determinados tipos de aplicações linguísticas, ao mesmo tempo em que possuem problemas que não são facilmente contornáveis, o que constitui o cerne do projeto de pesquisa do autor.