

Capítulo 5

Expressões multipalavras

Renata Ramisch
Carlos Ramisch
Aline Villavicencio

5.1 *Cenas dos próximos capítulos...*

O capítulo sobre as expressões multipalavras, que sairá na próxima edição deste livro, vai abordar um *tema indigesto* no universo no Processamento de Linguagem Natural (PLN). É que essas expressões ficam no limite entre a sintaxe e a semântica, e sempre acabam ficando *no meio do fogo cruzado*. De um lado, elas apresentam idiossincrasias e especificidades que não permitem determinadas operações sintáticas comuns a outros conjuntos de palavras. De outro, as definições do sentido das expressões multipalavras (ou MWEs) em geral fogem à regra de que o significado do todo se dá pela soma das partes. Ainda bem, pois imagine se a expressão “engolir o sapo” fosse literal.

Assim, a nossa intenção para este capítulo é, em primeiro lugar, definir quais são os conceitos fundamentais para quem vai *navegar pelas águas turbulentas* do tratamento computacional de MWEs. Entre esses conceitos, começaremos com a discussão sobre os elementos que compõem uma MWE: seriam palavras? Seriam lexemas?

Na sequência, abordaremos os conceitos de MWEs propriamente ditos. A literatura traz inúmeras maneiras de analisar essas expressões, e as definições são tão variadas quanto os campos de estudo que se interessam por esse tema. Em PLN, escolhemos a definição utilizada pelo projeto PARSEME (o qual também é apresentado brevemente no capítulo que está por vir):

Expressões multipalavras são entendidas como sequências (contínuas ou descontínuas) de palavras que (i) contêm pelo menos duas palavras componentes que são lexicalizadas, ou seja, sempre realizadas pelos mesmos lexemas, incluindo uma palavra principal e pelo menos uma outra palavra sintaticamente relacionada, e (ii) exibem algum grau de idiossincrasia lexical, morfológica, sintática e/ou semântica.

Considerando essa definição inicial de MWE, vamos pormenorizar alguns dos seus aspectos importantes, como idiossincrasias, coesão sintática e elementos lexicalizados. Ao mesmo tempo, também buscamos *separar o joio do trigo*, uma vez que, para entender o que é uma MWE, é fundamental que se saiba também o que ela não é. Então, explicaremos por que, na nossa abordagem, os compostos, as colocações e as metáforas não são expressões multipalavras.



Embora esta seja uma tarefa em aberto, apresentaremos, no nosso futuro capítulo, uma tentativa de classificação de MWEs. O grupo de expressões cujo núcleo é um verbo foi definido nas diretrizes de anotação do PARSEME, e nós basicamente utilizamos as definições desse projeto. No que se refere às demais categorias, consideramos uma série de aspectos práticos e teóricos na nossa proposta de categorização, levando em conta sobretudo a sua função em uma sentença. Traremos, ainda, algumas discussões sobre ambiguidade, variabilidade e arbitrariedade.

Após a etapa de conceituações, citaremos as principais tarefas de PLN que envolvem MWEs, assim como o estado da arte, os métodos e algoritmos que buscam resolver esse problema. Essas tarefas se dividem basicamente em dois grandes grupos: a) a descoberta/detecção; e b) a identificação de MWEs.

Você verá no capítulo, ainda, uma breve apresentação dos recursos existentes para o português brasileiro, assim como o conhecimento necessário para entender ou resolver alguma tarefa em torno da questão das MWEs. Depois, trataremos dos principais métodos de avaliação, assim como métricas e testes comumente utilizados pela comunidade para avaliar o desempenho de métodos e algoritmos direcionados a tratar ou resolver as tarefas antes mencionadas.

Por fim, consideramos importante *olhar pelo retrovisor* e entender qual foi o percurso para chegar até onde estamos, no tema das MWEs, em termos científicos, na direção da resolução dessa complexa tarefa. Em que ponto estamos hoje? Qual é a posição do português brasileiro em termos de recursos e de pesquisas diante da comunidade internacional? Quais são os desafios que permanecem aos pesquisadores e entusiastas das MWEs para *descascar esse abacaxi* em tempos de modelos de língua tão grandes que *não cabem em si*?

Sem querer *prometer mundos e fundos*, o capítulo que está por vir tentará acrescentar mais um tijolo na construção do conhecimento linguístico em PLN para o português brasileiro. Como você pôde perceber por este resumo, o capítulo contém uma série de exemplos de expressões multpalavras, tanto para ilustrar fenômenos linguísticos quanto para divertir leitores e leitoras. Vale a pena esperar pelas *cenias do nosso próximo capítulo*!

