Capítulo 10

Uma abordagem estatística para a identificação de colocações verbais usando o projeto AC/DC em www.linguateca.pt

Milena Uzeda Garrão e Maria Carmelita Padua Dias

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 10, p. 65–69, 2008

Tradicionalmente, pesquisas voltadas para o tratamento computacional de colocações vêm priorizando as combinações nominais ou os nomes compostos. Além de haver poucos estudos na área que se dediquem de forma sistemática às combinações verbais, existe um tipo em particular, o padrão V+SN, que se destaca das outras combinações verbais no português do Brasil (PB) e no europeu (PE) tanto pela sua freqüência quanto pelos seus alegados sub-padrões semânticos. O critério estatístico a partir de corpus adotado nesse projeto, como alternativa a uma abordagem baseada na intuição do pesquisador (Ranchhod, 2003), vem se mostrando altamente promissor no domínio das colocações do tipo V+SN (Uzeda Garrão, 2006; Uzeda Garrão e Dias, 2006).

Nossa metodologia, testada em padrões V+SN do PB pode ser resumida da seguinte forma: 1) uso de corpus etiquetado do PB como fonte de dados; 2) aplicação de um filtro para detecção de todos os padrões V+SN presentes no corpus; 3) aplicação de um teste estatístico ao filtro, chamado logaritmo de verossimilhança (Banerjee e Pedersen, 2003) para identificar as reais colocações (como "fazer parte", "tomar conta") em detrimento de combinações sintáticas casuais; 4) edição humana.

A justificativa para uma maior atenção descritiva voltada às colocações nominais em detrimento às colocações verbais se deve à importância do primeiro tipo de construção em textos de especialidade, auxiliando mais especificamente os domínios de Sumarização Automática e Recuperação de Informação. Acreditamos, contudo, que, embora se atribua à colocação verbal um papel secundário, ela vem a ser peça chave para o domínio de PLN. O padrão de combinação V+ SN ilustra claramente essa constatação, uma vez que inclui na sua estrutura um nome (SN), o que enfatiza a sua relevância também para domínios que priorizam o tratamento de colocações nominais.

10.1 Metodologia

10.1.1 O corpus utilizado: CETENFolha

O CETENFolha (Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo, `www.linguateca.pt/CETENFolha`) é um corpus jornalístico de cerca de 24 milhões de palavras em PB, parte integrante do corpus NILC (Pinheiro e Aluísio, 2003), ver também Aires e Aluísio (2001), que contém textos brasileiros do registro jornalístico, didático, epistolar e redações de alunos. Trata-se de uma parte de um corpus (Córpus NILC/São Carlos) com 37 milhões de palavras disponibilizado pelo projeto AC/DC (Santos e Sarmento, 2002).

De fato, a opção por um corpus de teor jornalístico tem suas implicações: a língua fica prioritariamente associada àquilo que é considerado notícia em detrimento, por exemplo, de uma conversa despretensiosa entre adolescentes. Entretanto, a escolha por esse tipo de extrato da língua também está associada à falta de um corpus mais robusto do PB. Uma outra razão da opção pelo corpus CETENFolha está no fato de, no ano de 2006, ser o único

significativo no PB disponível para *download*; e, portanto, o único passível de aplicação dos testes probabilísticos que virão mais adiante.

10.1.2 Aplicação do filtro para padrões V+SN aos verbos mais freqüentes

Com base nos 30 verbos com maior freqüência absoluta no corpus, partimos para uma restrição formal para obter os 10 verbos mais freqüentes seguidos facultativamente de determinante e obrigatoriamente de nome, formando a estrutura V+(det)+N. Tomando como exemplo o verbo *fazer*, o formalismo para tal detecção no AC/DC seria (**[lema="fazer"& pos="V"] [pos="DET.*"]? [pos="N"] [classe="JOCF"]**). A fórmula JOCF se refere à parte do corpus NILC/São Carlos que constitui o corpus CETENFolha. Obtivemos, finalmente, os 10 verbos mais freqüentes encabeçando uma estrutura V+SN. São eles: "fazer", "ter", "dar", "perder", "usar", "receber", "deixar", "tomar", "ganhar" e "criar".

Aplica-se então a todas as ocorrências desses 10 lemas no corpus, já baixado para a pesquisa, um filtro V+det+N, que foi viabilizado, nesse projeto, através de um programa, feito em linguagem Java, que recebe como entrada a ocorrência desses lemas no córpus e fornece como resultado a lista de todas as ocorrências de, por exemplo, fazer+(det)+ N (Nogueira, 2004). Somente na etapa seguinte é aplicado o teste estatístico e é estabelecida a lista das candidatas a colocações que, posteriormente, são ordenadas por freqüência.

10.1.3 A aplicação do logaritmo de verossimilhança aos padrões V+(det)+N encabeçados pelos verbos mais freqüentes no corpus

A aplicação do Logaritmo de Verossimilhança, foi disponibilizada através do pacote estatístico NSP (Banerjee e Pedersen, 2003). Após a sua aplicação é estabelecida a lista das candidatas a colocações que, posteriormente, são ordenadas por freqüência. Dentre os métodos de Testagem de Hipótese fornecidos em NSP, o Logaritmo de Verossimilhança tem por objetivo detectar se um bigrama é mais do que uma simples co-ocorrência casual na língua. Esse tipo de testagem requer a formulação de dois tipos de hipóteses formalizadas abaixo:

$$H_1 : P(w_1 \mid w_2) = P(w_1 \mid \neg w_2)$$
$$H_2 : P(w_1 \mid w_2) \neq P(w_1 \mid \neg w_2)$$

Onde H = hipótese, P = probabilidade, w = palavra

Por exemplo, assumindo que a expressão *fazer sucesso* seja uma colocação, espera-se que a hipótese de dependência H_2 : $P($ fazer $\mid$ sucesso $)$ $\neq$ $P($ fazer $\mid$ $\neg$ sucesso $)$ seja

verdadeira e que a hipótese de independência $H_1 : P(\text{ fazer} \mid \text{sucesso }) = P(\text{ fazer} \mid \neg \text{ sucesso })$ seja falsa. Portanto, o método avalia a probabilidade de H_2 ocorrer em detrimento de H_1.

10.1.4 Resultados e Edição Humana

Sob uma perspectiva quantitativa o método se revelou satisfatório. Em outras palavras, dentre as 1000 candidatas a colocações apontadas pelo método (100 de cada um dos verbos listados na seção 10.1.2, apenas 128 foram consideradas ruído. Um acerto de 87,2%. As "pseudo-colocações" extraídas pelo método, ou seja, os "deslizes" por ele cometido (12,8%), foram indicadas na listagem final da seguinte forma:

1. Erro de avaliação estrutural. Este tipo de erro pode ter sido cometido pelo método por duas razões principais: em função da etiquetagem equivocada no corpus; em função de o método ter considerado uma janela sintática menor do que a expressão representa (JAN): *ter um papel*, por exemplo, foi detectado pelo método como uma colocação do padrão procurado quando, na verdade, sua estrutura vai além de V+(det)+N.

2. Outros ruídos foram atribuídos exclusivamente ao corpus: colocações claramente datadas: como *criar a URV*, *usar a URV*, *tomar AZT*.

Há outros dois tipos de interferência na detecção de colocações que não foram considerados propriamente ruídos. São eles: recursos coesivos, como a utilização de anáfora: alguns exemplos são *fazer a denúncia*, *dar a notícia*, *ter a doença* (COE) e omissões de artigo (tanto definido quanto indefinido), características de manchetes de jornal, como Presidente da Shell ***deixa cargo*** amanhã.

A título de ilustração, a tabela 10.1 diz respeito às 100 colocações do tipo Fazer+(det)+SN mais freqüentes detectadas no corpus. Na verdade, a listagem segue até que se chegue a co-ocorrências menos freqüentes. Esse é apenas um pequeno extrato do que o teste foi capaz de gerar. Tomemos como exemplo o primeiro bigrama da lista, *fazer parte*. O número que segue à colocação (1) diz respeito à sua posição em relação às outras colocações. O segundo número (2805) se refere ao número de ocorrências no corpus.

10.2 Conclusões e trabalhos futuros

A grande vantagem deste método está no seu teor preditivo. Através dele, podemos constatar preferências de usos das expressões presentes no corpus. Portanto, o que consideramos especialmente relevante nesta abordagem com base em corpus, é que não fazemos conjecturas daquilo que ocorre e não ocorre em uma língua, pois uma perspectiva exclusivamente intuitiva pode ser muitas vezes contra-argumentada por dados reais da língua.

fazer parte,1,2805	fazer falta,17,124	fazer exames,33,86
fazer campanha,2,616	fazer acordo,18,123	fazer um discurso,34,83
fazer questão,3,485	fazer alguma coisa,19,112	fazer referência,35,82
fazer sucesso,4,289	fazer propaganda,20,112	fazer um teste,36,82
fazer compras,5,227	fazer o gol,21,107 (COE)	fazer uma avaliação,37,76
fazer papel,6,189 (JAN)	fazer greve,22,105	fazer um balanço,38,76
fazer sentido,7,177	fazer perguntas,23,104	fazer gols,39,75
fazer comício,8,176	fazer sua estréia,24,98	fazer o teste,40,74 (COE)
fazer um filme,9,151	fazer uso,25,98	fazer o pedido,41,74 (COE)
fazer um acordo,10,149	fazer filmes,26,95	fazer shows,42,74
fazer a conversão,11,143 (COE)	fazer coisas,27,95	fazer palestra,43,72
fazer um trabalho,12,142	fazer testes,28,95	fazer a ligação,44,71 (COE)
fazer mal,13,130 (JAN)	fazer uma campanha,29,92	fazer a festa,45,68
fazer sexo,14,130	fazer oposição,30,90	fazer as contas,46,68
fazer política,15,126	fazer exercícios,31,89	fazer uma pesquisa,47,67
fazer críticas,16,125	fazer um levantamento,32,88	fazer muito tempo,95,45 (JAN)

Tabela 10.1: Colocações Fazer+(det)+SN mais freqüentes no corpus.

Nosso olhar eminentemente empírico é capaz de detectar preferências de usos ao invés de intuir aquilo que pode ou não ocorrer em um corpus, com base em testes de aceitabilidade, comumente utilizados para identificar as colocações.

Em suma, acreditamos que esse trabalho tenha uma função prática e teórica para a lexicografia. Sua natureza genuinamente estatística viabiliza uma rápida construção de uma base de dados robusta das colocações verbais mais freqüentes através de evidência empírica. Uma das próximas etapas é utilizar uma ferramenta mais recente desenvolvida para identificação de colocações - Linguistics Tool (Caminada, 2008) - e comparar os resultados. Pretendemos também estender o método para detectar outros padrões de colocações freqüentes encabeçadas por verbo (ex.: V+Prep+N; V+N+prep+N) e contribuir de forma efetiva para a lexicografia computacional do PB e, futuramente, para a lexicografia do PE, trabalhando em conjunto com pesquisadores nativos da modalidade européia.