Matrizes fraseológicas em artigos de Medicina: um estudo com vistas ao desenvolvimento de ferramenta automática de apoio à tradução

Viviane Possamai, Maria José B. Finatto

Programa de Pós-Graduação em Letras – Universidade Federal do Rio Grande do Sul
91501-970 – Porto Alegre – RS – Brasil

vivianepossamai@yahoo.com.br, mfinatto@terra.com.br

Abstract. *The current study presents an overview of a further study that has been carried out about phraseological matrices and their applicability to a wider scope – a doctoral thesis – whose objective is to develop heuristics for the construction of a system that firstly automatically recognizes relevant information to translators and, secondly, provides translators with possible suggestions of translations already published over the Web. To that end, we have been using Corpus Linguistics tools.*

Resumo. *O objetivo deste trabalho é demonstrar parte de um estudo realizado sobre matrizes fraseológicas e sua aplicabilidade em um âmbito maior, um estudo de doutorado, que visa o desenvolvimento de heurísticas para a construção de um sistema de reconhecimento automático de informações relevantes para tradutores de artigos científicos e, posteriormente, possíveis equivalentes. Para este fim, utilizamos ferramentas da Lingüística de Corpus.*

1. Introdução

Pesquisadores brasileiros têm se empenhado no esforço de terem sua produção intelectual divulgada na forma de artigos científicos, em sua maioria, em publicações do tipo *journal* ou anais de eventos. Nesse contexto, publicações em periódicos de qualidade e internacionais são um trunfo, tanto pelo alcance da divulgação das pesquisas como pelo peso que têm em avaliações de agências de fomento.

Nesse cenário, estudantes e pesquisadores acabam encontrando como obstáculos para a realização profissional, não a qualidade de seus estudos, mas a dificuldade de redigir um texto em sua língua materna e, para muitos ainda, em língua estrangeira, especialmente a língua inglesa.

Todo esse cenário acaba favorecendo o mercado de trabalho de profissionais que trabalham com o texto escrito, como revisores e tradutores, ainda que, muitas vezes, estes também não tenham tido treinamento para trabalhar com o texto acadêmico-científico e nem encontrem oportunidades de qualificação. Com esta preocupação é que estamos realizando um estudo de doutorado que tem como objeto o texto do artigo científico e, como objetivo prático final, a elaboração de aportes para a construção de um sistema automático que, num primeiro momento, reconheça informações relevantes para a tradução e, subseqüentemente, busque num banco de dados e/ou Internet, possíveis equivalentes para as informações reconhecidas. Como objetivo teórico, queremos defender uma tese sobre estruturas textuais (ou seqüências de caracteres) que

carreguem um valor importante dentro do processo tradutório – pela dificuldade de tradução, pela especificidade do tema e pela prototipicidade de uso – sugerindo também que o texto pode ser considerado um objeto de estudos em Terminologia.

2. Objetivo

O objetivo do presente trabalho é verificar se matrizes fraseológicas, conforme delineado por Gouadec [1994], podem nos ajudar a desenvolver um conhecimento que possa ser aplicado em uma ferramenta automática, que vem a ser a porção prática do estudo. Contamos com a hipótese de que ao identificar as matrizes, juntamente com as palavras-chave que as compõem, teremos acesso a palavras que aglutinam em torno de si construções textuais e informações que são típicas dos textos em questão. Pretendemos identificar blocos textuais em que determinadas palavras-chave, típicas dos artigos médicos, são encontradas (ex.: *estatisticamente, significante, evidência, corte transversal etc*), em estruturas como: *a presença de x é estatisticamente significante em, não houve evidência de modificações, em um estudo de corte transversal etc*).

Segundo Gouadec [1994], matrizes são cadeias notáveis de caracteres, nas quais o caráter de notabilidade se dá por uma ou mais de quatro razões: *caráter especializado, repetição, risco inerente a sua manipulação e vantagem que se pode ter ao dominá-las*. O autor trabalha com uma idéia bastante ampla de cadeias de caracteres especializadas. Essas cadeias têm tamanho variável, incluindo palavras, grupos de palavras, termos, locuções, expressões, orações, segmentos de frases, frases, conjuntos de frases. Essas matrizes contam com uma parte fixa (ex.: O objetivo deste trabalho é [X]) e uma parte variável (ex.: X, na sentença anterior). Cada uma dessas partes contém palavras-chave. Neste trabalho, apresentamos apenas uma amostra de nosso estudo de observação de matrizes em artigos médicos, dadas as limitações e o caráter inicial da pesquisa.

3. Materiais e métodos

O texto escolhido como ponto de partida para este estudo é o resumo de um artigo que se intitula *Prevalência de rinite alérgica em adolescentes do Distrito Federal: comparação entre as fases I e III do ISAAC*, publicado no periódico *Jornal de Pediatria*, volume 82(2), no ano de 2006. Para projetar os resultados encontrados da observação deste artigo, utilizamos um córpus constituído de 231 artigos do mesmo *Jornal de Pediatria* (496.145 tokens e 21.502 types) [Coulthard, 2005].

O primeiro passo realizado foi a identificação manual das seqüências notáveis, segundo os critérios definidos por Gouadec. Reconhecemos oito seqüências que consideramos serem possíveis matrizes. Na seção a seguir, apresentamos, a título de exemplo, apenas uma dessas seqüências.

Após o reconhecimento, destacamos o que consideramos as palavras-chaves da seqüência. Esse procedimento tem fundamento em uma noção também explicitada por Gouadec, de que “é preciso... que as chaves sejam aquelas que espontaneamente os usuários do fichário mobilizam". Selecionamos palavras que têm potencial de agregarem diferentes tipos de construções ao seu redor, sem diferenciar categoria gramatical, nem diferenciar termo de palavra.

O segundo passo foi a confirmação de que a seqüência selecionada tinha características de matriz, isto é, suas variáveis podem ser preenchidas por outras palavras ou seqüências. Neste passo, utilizamos o córpus mencionado.

Com o software WordSmith Tools e sua ferramenta Concord fizemos uma busca nos textos visando encontrar outras construções com a mesma palavra-chave das seqüências selecionadas, ou, dizendo de outra forma, fizemos uma busca das concordâncias das palavras-chave selecionadas.

4. Resultados

A título de exemplo, apresentamos apenas uma, das oito seqüências de caracteres extraídas, e alguns comentários a respeito do que foi possível observar.

> **...determinar a prevalência de** [rinite alérgica] **em** [um grupo aleatório de escolares de Brasília (DF)], [com idade entre 13-14 anos...]

No segundo passo, trabalhamos com a seguinte notação matricial

...<u>determinar a prevalência</u> de [X] em [Y], [Z]

X – evento (sintoma, doença)
Y – população/sujeitos
Z – complemento/especificação de Y, especialmente um dado de idade

Sete foram os resultados de concordâncias com a seqüência “determinar a prevalência de [x] em [y]”, no córpus de 231 artigos.

A matriz se mostrou produtiva, uma vez que as variáveis pressupostas foram preenchidas por elementos condizentes com o pressuposto (X, Y, Z; sintoma, sujeitos e complemento, respectivamente) em todas as ocorrências. Os itens que preencheram as variáveis X, Y e Z são apresentados a seguir (obs.: consideramos que a variável Z não precisava ser sempre preenchida):

a) EA e sintomas relacionados, estudantes da região centro-sul

b) colonização pelo S. pneumoniae, crianças que freqüentam creches municipais de Taubaté

c) tabagismo, adolescentes argentinos, entre os 11 e os 15 anos de idade

d) soropositividade para toxoplasmose, gestantes

e) sobrepeso e obesidade, uma amostra de crianças da rede de ensino público e privado da zona urbana de Feira de Santana-BA

f) bactérias anteriormente referidas, efusão da orelha média de crianças com OME

g) aleitamento, crianças que nunca tiveram diarréia

Chamou a atenção nessa etapa o fato dessa matriz ocorrer com a palavra *objetivo* em cinco das 8 ocorrências. Pretendemos avaliar mais profundamente o resultado desse dado no decorrer do trabalho, uma vez que para os propósitos da tese pode revelar-se um dado importante, pois pretendemos também trabalhar com relacionamentos estatísticos entre palavras**.**

Não menos importante é a revelação de que a variável X configura-se uma seqüência de caracteres com um caráter mais aproximado ao de termo canônico, sendo composta evidentemente de conceitos que assumem um caráter terminológico no contexto em que se encontram. Assim, podemos dizer que se um sistema automático tivesse em sua base de dados a matriz **determinar a prevalência de [X] em [Y]** como recurso para extração automática de candidatos a termo**,** pelos dados levantados, haveria 100% de acerto que X seria um termo da área médica.

5. Conclusão

Dentro dos objetivos a que estamos nos propondo, e considerando que o trabalho descrito acima constituiu-se apenas de uma investigação preliminar, acreditamos que já temos alguns indícios de que uma abordagem de matrizes pode revelar-se produtiva. Construindo uma situação hipotética, em que um sistema de reconhecimento automático consideraria que as palavras **determinar** e **prevalência** tivessem um índice de co-ocorrência, o sistema poderia fazer uma busca na Web, usando as palavras em inglês, *determine* e *prevalence of*, mais um dos termos presentes no título do artigo, que já estariam traduzidos em um dicionário, teríamos a seqüência:

> *This study was undertaken to **determine** the **prevalence** of asthma, eczema, and **allergic rhinitis** in school children in Kota Bharu*

que, consideramos, ajudaria o tradutor a construir uma seqüência semelhante para seu texto, porém mais criativa, talvez, do que a que ele alcançaria fazendo a tradução literal *The objective of this study was to determine the prevalence of…*

As palavras-chave selecionadas neste estudo, em sua maioria, fazem parte de estruturas maiores, as matrizes, e replicam-se em um universo maior de textos. Esse caráter de repetição justifica a notabilidade e a especificidades dessas estruturas, tanto as matrizes como as palavras-chave.

Ainda estamos no início das observações, mas, após essa investigação, que visou principalmente manusear as informações, observar, ver como se comportam, confirmamos a dificuldade que existe em se sistematizar resultados levando em conta todos os fatores lingüísticos envolvidos. Isso talvez esteja indicando um outro caminho para a obtenção dos resultados que pretendemos, como o uso de métodos mais estatísticos ou matemáticos, o que pode dar maior robustez à ferramenta. Para tal, realizaremos extensa revisão bibliográfica da área computacional que trata de classificação de textos, extração de informações e cálculos estatísticos de co-ocorrências de palavras. Avaliações quanto à validade da proposta serão realizadas com tradutores aprendizes, tendo em vista que esses seriam os principais beneficiados por uma ferramenta que identificasse padrões textuais típicos dos artigos médicos.

6. Referências

1. GOUADEC, Daniel. Nature et traitement des entités phraséologiques. Terminologie et phraséologie. Acteurs et amenageurs. Actes du deuxième Université d'Automne en Terminologie. Paris: La Maison du Dictionnaire, 1994. p. 164-193.

2. COULTHARD, R. James. The application of Corpus Methodology to Translation: the JPED parallel corpus and the Pediatrics comparable corpus. Dissertação de Mestrado, 2005, 155 p.