Linguistics Tools – Uma Plataforma Expansível de Funções de Consulta a Corpus

Nuno Caminada
Instituto Militar de Engenharia
Praça General Tibúrcio, 80,
Praia Vermelha, Urca.
Rio de Janeiro, Brasil
55.21.2546.7080
nunocaminada@gmail.com

Violeta Quental
Pontifícia Universidade Católica
do Rio de Janeiro
Rua Marquês de São Vicente
225, Gávea.
Rio de Janeiro, Brasil
55.21.3527.1001
violetaq@puc-rio.br

Milena Garrão
Pontifícia Universidade Católica
do Rio de Janeiro
Rua Marquês de São Vicente
225, Gávea.
Rio de Janeiro, Brasil
55.21.3527.1001
migarrão@terra.com.br

RESUMO

Este artigo descreve a ferramenta de consulta a corpus Linguistics Tools, projetada para a identificação de multivocábulos preposicionados em corpora da língua portuguesa, utilizando algoritmos clássicos como o Students T-Test, o Log Likelihood e o Mutual Information, mas permitindo sua extensão tanto para funções de tratamento de corpus quanto de algoritmos de identificação. Esta ferramenta foi desenvolvida na linguagem Java e aceita como entrada corpora anotados morfossintaticamente pelo parser PALAVRAS [2]. Uma descrição da ferramenta será feita, bem como uma comparação de resultados sobre dois corpora, de tamanhos similares, mas com características diferentes.

ABSTRACT

This paper describes Linguistics Tools, an extensible corpus query tool designed for the search for prepositional multi-word expressions in corpora of the Portuguese language, using classic algorithms such as T-Test, Log Likelihood and Mutual Information, but also leaving room for the implementation of further parsing and identification functions and algorithms. This tool was developed in the Java language and takes as input corpora annotated by the parser PALAVRAS (Bick2000). A description of the tool is given, and results from two corpora of different characteristics but of the same size are presented and compared.

Categories and Subject descriptors

J.5[Arts and Humanities] Linguistics.
D.2.13[Reusable Software] Reusable Libraries.

General Terms

Design.

Keywords

Lingüística, Lexicografia computacional multivocábulos, expressões cristalizadas, collocations, corpus, Engenharia de Software.

1. INTRODUÇÃO

A identificação de expressões multivocabulares, termo que abarca expressões com diferentes graus de fixidez de composição e opacidade semântica, é uma das tarefas lexicográficas que mais se beneficiam de métodos e aplicações computacionais. A identificação de multivocábulos tem caráter estatístico, visto que o que distingue este tipo de expressão é exatamente sua prevalência frente às outras combinações de vocábulos da língua.

As definições de multivocábulos variam de autor para autor, assim como os critérios utilizados para sua identificação. Uma contribuição importante da lexicografia computacional nesta área tem sido a aplicação em corpora de testes matemáticos e estatísticos que indicam frequência relevante para a caracterização de multivocábulos. Desloca-se a questão com isso para o uso estatisticamente consistente de uma determinada expressão, e não apenas para suas características lingüísticas.

Seguindo essa tendência, apresentamos aqui uma ferramenta extensível para a busca e identificação de bigramas e trigramas multivocabulares da língua portuguesa, baseados em padrões gramaticais definidos pelo operador em tempo de execução, tendo como entrada corpora anotados da língua portuguesa submetidos a um processo de atomização também descrito adiante, que quebra os multivocábulos previamente identificados pelo parser PALAVRAS e permite uma identificação em tabula rasa, que ajuda no processo de validação destes multivocábulos e no processo de identificação de novas expressões.

Em seu estado atual, esta ferramenta implementa cinco algoritmos largamente descritos e utilizados na literatura de multivocábulos e colocações: o Teste-T, o *Chi-Square*, o *Log Likelihood*, o *Mutual Information* e o *Dice Coefficient* [7]. Cada um destes algoritmos pode ser aplicado a diferentes corpora e diferentes padrões gramaticais, permitindo ao operador comparar resultados entre padrões e tipos de corpora, assim como a maximização do processo de identificação.

2. A FERRAMENTA *LINGUISTICS TOOLS*

A ferramenta *Linguistics Tools* foi projetada para executar duas funções principais, a de tratamento (*parsing*) de corpora e a de

análise de multivocábulos. Para o *parsing*, a ferramenta possui uma classe extensível (classe **parser**) que possui duas versões do método abstrato **parseLine** e o método concreto **write**. A criação de serviços de processamento de texto é realizada através da extensão desta classe. As classes estendidas implementam suas próprias interfaces gráficas, concentrando assim nestas classes todas as alterações necessárias à implementação de novos serviços. A baixa coesão caracterizada pela sobreposição de atribuições, neste caso interface gráfica e adição de novos algoritmos de identificação.

A Figura 1 mostra o diagrama de classes da aplicação, com serviços de *parsing* e algoritmos de identificação de multivocábulos instanciados.

A instanciação de serviços de *parsing* é feita através de uma classe fabricante de *parsers*, **ParseFactory,** que, como o nome sugere, implementa o padrão de projeto[3] Factory Method [3], que auxilia na coesão da classe de entrada da aplicação,

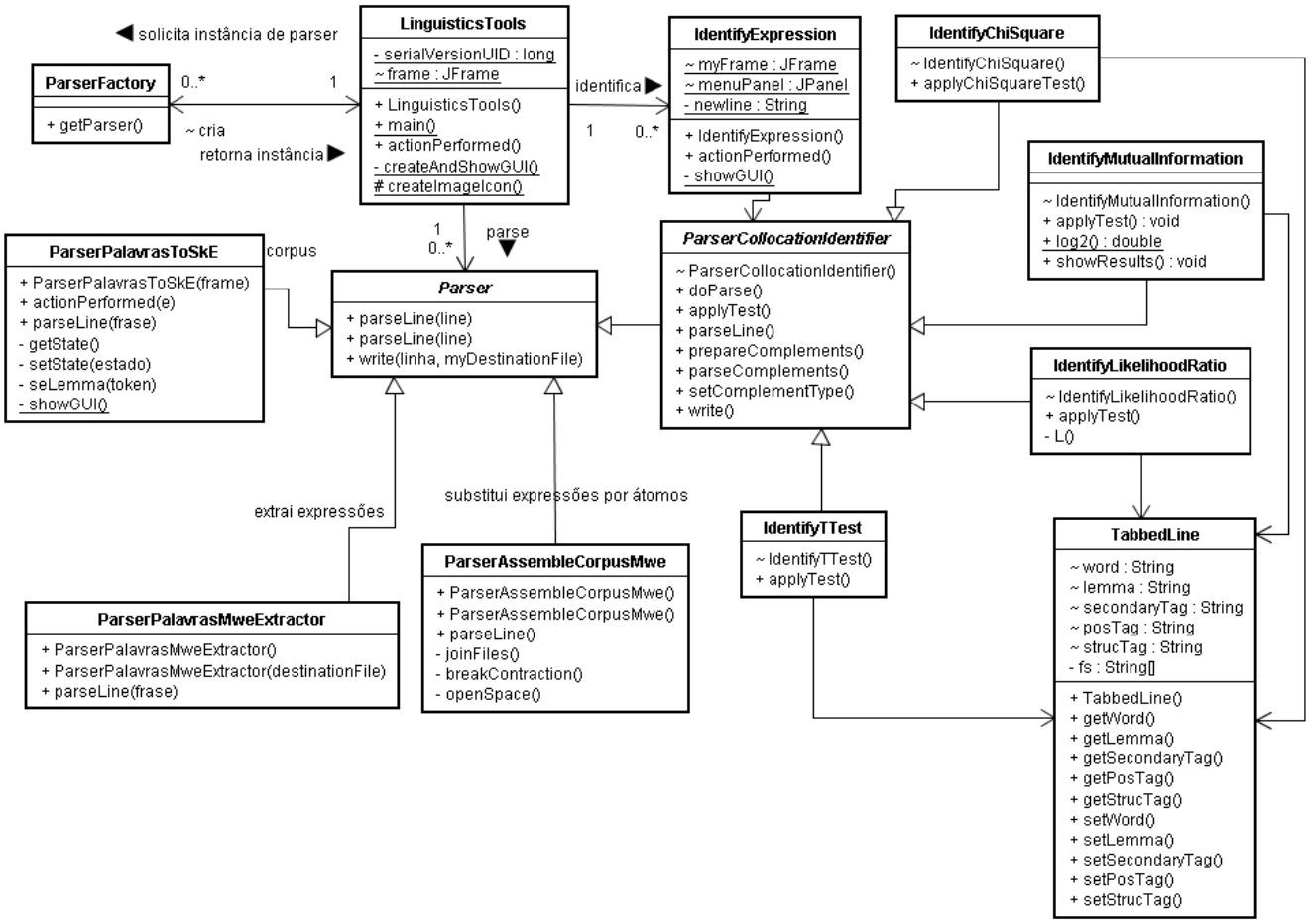

Figura 1 – Diagrama de classes da aplicação

lógica, é compensada pela total encapsulação do serviço, que simplifica a manutenção e o funcionamento geral da aplicação.

Para os serviços de análise e identificação, a classe **ParserCollocationIdentifier** também pode ser estendida, sendo possível assim implementar diferentes algoritmos de identificação e diferentes versões destes algoritmos. Como na extensão da classe **parser**, a classe que estende acumula atribuições gráficas e lógicas, perdendo em coesão[1] mas ganhando em encapsulamento[2], facilitando assim o processo de separando estas das instâncias de *parsers*. Desta maneira, para adicionar um novo *parser* é necessário somente alterar a *factory* para incluir o novo serviço, e criar o *parser* com as capacidades desejadas.

3. CORPORA

Os resultados aqui apresentados foram obtidos a partir da aplicação do método de identificação sobre dois corpora bem distintos.

O primeiro corpus, denominado aqui **Corpus Jornalístico**, é

[1] Coesão é definida como a propriedade de uma classe de possuir um único propósito ou função, não devendo ser subdividida em classes separadas.

[2] Encapsulamento é definido como a ocultação dos processos internos das classes, expondo ao usuário apenas o necessário para o funcionamento do sistema.

[3] Padrões de projeto são padrões de desenvolvimento de software que foram catalogados por [3] devido a sua importância estrutural e/ou por facilitar o reuso. Sua utilização é uma das bases do desenvolvimento de aplicações orientadas a objeto.

composto de textos de um jornal de grande circulação e possui material contemporâneo brasileiro, com não mais que uma década de idade, somando ao todo 32.044.437 tokens, que podem ser palavras, sinais de pontuação etc.

O segundo corpus, denominado **Corpus Internet,** foi construído totalmente a partir da ferramenta WebBootCat [5], que realiza coleta de textos na internet a partir de parâmetros definidos pelo utilizador, como listas de palavras sementes. Estas palavras são aplicadas à API do Yahoo!, gerando uma coleção de textos cujo tamanho varia de acordo com os parâmetros, mas variam entre algumas centenas de milhares de palavras e alguns milhões de palavras.

No caso específico dos textos coletados para esta pesquisa, estas palavras sementes foram extraídas do **Corpus Jornalístico**, a partir de uma seleção dos 500 substantivos comuns com maior ocorrência. Destes 500, 100 foram separados em 10 grupos de 10 palavras, que foram submetidos ao WebBootCat, gerando corpora de 3 milhões de palavras em média. É importante ressaltar que a definição de corpora utilizada aqui é aquela proposta por Kilgarriff [4]: "Um corpus é uma coleção de textos quando considerado como um objeto de estudo literário ou da língua[4]".

Ambos os corpora foram anotados morfossintaticamente pela ferramenta PALAVRAS [2] e sofreram um processo de atomização para reverter as expressões multivocabulares já identificadas por esta ferramenta e assim poder ter seus multivocábulos re-identificados e avaliados contra a evidência de corpus.

4. IDENTIFICAÇÃO

O processo de identificação é baseado na localização de preposições no corpus a partir de sua anotação morfossintática.

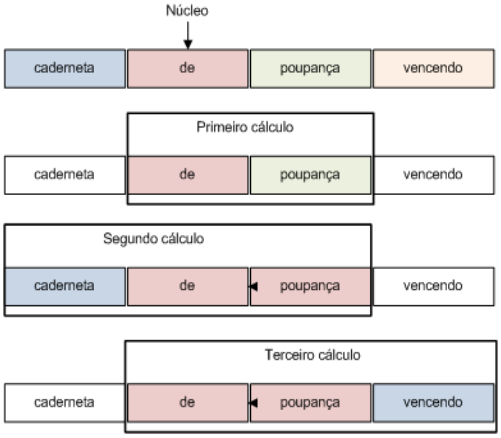

Figura 2. Janelas deslizantes

Quando uma preposição é identificada, a palavra anterior e as duas subseqüentes são armazenadas. Em uma varredura subseqüente, a aplicação obtém a quantidade de ocorrências de cada uma das palavras e a quantidade de ocorrências da preposição combinada com estas palavras. Estes dados, mais o número total de palavras no corpus, são os dados necessários à aplicação da classe de algoritmos suportadas por este *framework*, que inclui, além dos algoritmos citados acima, outros como o Z-Test.

Com estes quatro vocábulos, é possível a composição de uma janela lógica de bigramas e trigramas, conforme apresentado na Figura 2, envolvendo a preposição e a palavra seguinte (PREP+*), a palavra anterior, a preposição e a palavra seguinte (*+PREP+*) e a preposição e as duas palavras seguintes (PREP+*+*).

Para a composição de resultados, a janela do segundo cálculo foi utilizada para aferir se o bigrama identificado no primeiro cálculo era parte de uma expressão maior, assim como os trigramas do terceiro cálculo. Como resultado, somente bigramas e trigramas que não demonstraram propensão a participar de expressões maiores, como "informal de variação", que obteve score 6.7 no Teste-T mas que faz parte de expressões maiores como "faixa informal de variação cambial" ou "banda informal de variação de câmbio", foram apresentados como resultados do trabalho. Porém, a janela do segundo cálculo apresentou resultados interessantes para a identificação de expressões multivocabulares, especialmente para o padrão N+PREP+N, e por isso serão apresentados aqui também.

O processamento preferencial de bigramas e trigramas utilizando a metodologia de janelas permite o rápido processamento de corpora com mais de 30 milhões de palavras, sem exceder os limites de memória de computadores de linha, comumente utilizados por lingüistas em tarefas como esta. Como parâmetro, o processamento de um corpus com 32 milhões de palavras num computador com processador Pentium Dual Core e 2Gb de RAM levou aproximadamente sete minutos para gerar resultados finais.

5. RESULTADOS

As Tabelas 1 e 2 apresentam os 13 resultados de maior classificação no Teste-T da identificação de multivocabulos preposicionados do Corpus Jornalístico. O score pré-janela denota a classificação antes da fatoração da janela do segundo cálculo, onde o bigrama é avaliado para determinar se faz parte de uma expressão maior.

A mesma listagem é apresentada no Score subtraindo janela, agora com a fatoração. É interessante notar que todos os bigramas apresentados na primeira coluna foram identificados como parte de expressões maiores e foram descartados na classificação para a segunda coluna.

A primeira coluna da Tabela 2 apresenta os 13 trigramas de maior classificação com núcleo preposicional. Estes trigramas foram identificados a partir do bigrama composto pela preposição, com padrão PREP+N, e do vocábulo anterior, que, nestes casos de maior classificação, é preponderantemente um N.

Tanto para a primeira quanto para a segunda coluna da Tabela 2, o problema da identificação foi reduzido a dois problemas binários, o primeiro envolvendo o bigrama PREP+N, e o segundo considerando este bigrama como um único vocábulo e realizando seu cálculo com o vocábulo anterior (Tabela 2, primeira coluna) ou posterior (Tabela 2, segunda coluna).

4 A corpus is a collection of texts when considered as an object of language or literary study .

Tabela 1. Classificação Teste-T, Corpus Jornalístico, (*+) PREP+N (+*), antes da aplicação da janela e com a subtração da janela

Score pré-janela		Score subtraindo janela	
Bigrama	**Score**	**Bigrama**	**Score**
por;exemplo	88,841	como;forma	18,475
em;relação	75,071	segundo;informação	15,179
por;causa	64,957	em;suma	14,607
em;torno	50,49	como;parte	13,91
a;favor	49,073	em;detrimento	13,318
de;juro	46,943	em;mina	13,093
de;segurança	43,189	por;motivo	12,915
de;saúde	42,842	para;uso	12,861
de;forma	40,546	sob;pena	12,514
em;média	39,269	com;efeito	11,793
de;futebol	39,26	segundo;especialista	11,728
por;volta	38,978	como;candidato	11,703
em;geral	37,827	em;protesto	11,702

Tabela 2. Classificação Teste-T, Corpus Jornalístico, (*+) PREP+N (+*), com as janelas do segundo e terceiro cálculo

Janela do segundo cálculo		Janela do Terceiro cálculo	
Trigrama	**Score**	**Trigrama**	**Score**
taxa;de;juro	45,134	em;relação;a,	79,164
assessoria;de;imprensa	31,028	por;causa;de	60,895
final;de;semana	30,768	em;torno;de	48,311
projeto;de;lei	30,46	em;vez;de	39,146
milhão;de;pessoa	29,905	por;volta;de	37,225
%;em;relação	29,122	com;base;em	36,125
fundo;de;pensão	24,386	por;parte;de	31,731
ano;de;idade	23,669	em;frente;a	31,603
plano;de;saúde	22,224	em;nome;de	31,055
cartão;de;crédito	21,925	em;busca;de	30,342
hoje;em;dia	21,655	a;respeito;de	29,763
milhão;de;tonelada	21,488	com;agência;internacional	29,705
ano;de;prisão	21,453	a;favor;de	28,814

Na segunda coluna da Tabela 2, a preponderância é de locuções prepositivas, o que não é surpresa dada sua prevalência quantitativa sobre os diferentes tipos de multivocábulos preposicionados. O viés do corpus jornalístico é aparente na expressão “com agência internacional” na quarta coluna, bem como em “milhão de pessoa” e “milhão de tonelada” na terceira coluna.

O importante é notar a alta taxa de precisão na identificação dos multivocábulos pesquisados, lembrando que os resultados denotam as formas lematizadas dos vocábulos, conforme anotados pela ferramenta PALAVRAS.

As tabelas 3 e 4 apresentam os 13 resultados de maior classificação para o Corpus Internet, seguindo a mesma metodologia empregada para a obtenção das Tabelas 1 e 2, ou seja, com o emprego do Teste-T e com a mesmo método de fatoração em janelas.

Tabela 3. Classificação Teste-T, Corpus Internet, (*+) PREP+N (+*),antes da aplicação da janela e com a subtração da janela

Score pré-janela		Score subtraindo janela	
Bigrama	**Score**	**Bigrama**	**Score**
até;hoje	33,266	como;sempre	13,678
até;agora	33,207	há;quase	10,983
de;aí	32,874	até;aí	10,468
por;aí	27,516	sem;nunca	9,407
para;trás	21,55	há;pouco	8,22
para;baixo	20,577	desde;cedo	8,147
por;lá	19,602	como;algo	7,64
até;aqui	18,048	desde;ontem	7,294
para;cá	16,8	sem;sequer	7,277
de;hoje	16,799	durante;quase	7,208
até;onde	16,071	até;ali	6,651
sobre;como	16,041	pra;onde	6,416
até;lá	15,777	sem;jamais	5,809

Tabela 4. Classificação Teste-T, Corpus Internet, (*+) PREP+N (+*), com as janelas do segundo e terceiro cálculo

Janela do segundo cálculo		Janela do Terceiro cálculo	
Trigrama	**Score**	**Trigrama**	**Score**
dia;de;hoje	18,277	a;não;ser	24,92
e;de;aí	14,515	de;aqui;a	17,565
cuidado;para;não	11,256	por;não;ter	16,63
sair;de;lá	10,611	de;hoje;-	16,365
parte;de;baixo	10,079	de;aí;que	15,259
sair;de;aqui	9,337	de;onde;vir	14,788
exemplo;de;como	9,144	de;não;ter	14,421
fato;de;não	9,073	por;aí;-	12,886
que;até;hoje	8,935	para;trás;-	12,577
olhar;para;trás	8,649	até;hoje;-	12,118
idéia;de;como	8,573	até;agora;-	12,029
saber;de;onde	8,519	de;mais;um	11,018
ir;para;lá	8,287	de;sempre;-	10,924

É interessante notar a variação nos multivocábulos encontrados, indicando que a variação de corpora altera significantemente a prevalência de multivocábulos. Na comparação direta fica ainda mais evidente o viés do Corpus Jornalístico, através de trigramas identificados na janela do terceiro cálculo como “taxa de juros” e “assessoria de imprensa”, que no Corpus Internet são substituídos em freqüência pelos menos formais “dia de hoje” e “e de+aí”.

6. TRABALHOS RELACIONADOS

Este trabalho relaciona-se fortemente com [8], e em menor

escala com [6].

No primeiro, os autores dão ênfase à caracterização lingüística do fenômeno das expressões multivocabulares na língua portuguesa, em suas diversas formas, e por fim utilizam o sistema Córtex [1] para sua identificação em corpora.

No segundo, os autores exploram as diferenças entre corpora da língua alemã, um contendo informações jornalísticas e o outro contendo textos extraídos de grupos de conversa da Internet, comparando os resultados através da aplicação dos mesmos algoritmos utilizados neste trabalho, o que permite um paralelo interessante com o trabalho aqui descrito.

7. CONCLUSÕES E TRABALHOS FUTUROS

A facilidade de uso e a rapidez na obtenção de resultados forneceram a lingüistas e lexicógrafos uma plataforma versátil de identificação de multivocábulos na língua portuguesa. Os resultados foram satisfatórios e permitiram a compilação de uma lista de sugestões de adição ao léxico do PALAVRAS que ainda está sendo revista. Além disso, a geração de grandes listas candidatas à classificação como expressões multivocabulares preposicionadas abriu novas frentes de trabalho para a determinação de parâmetros de corte para os diferentes algoritmos utilizados, o que terá impacto em todas as aplicações lingüísticas destes algoritmos.

Como ponto de partida foi fixada a preposição como núcleo do multivocábulo, mas nada impede que qualquer outra classe gramatical ou morfema seja empregada para este fim. Desta forma, uma das primeiras tarefas na pesquisa futura é a ampliação do escopo de possibilidades de núcleo do multivocábulo, assim como a avaliação dos impactos desta mudança no consumo de recursos computacionais, visto que ao longo da execução a máquina virtual Java mantém grande parte das relações identificadas sob a forma de *hashtables* em memória.

8.REFERÊNCIAS BIBLIOGRÁFICAS

[1] ARANHA, CHRISTIAN**.** Uma Abordagem de Pré-processamento para Mineração de Textos em Português. Tese de Doutorado, PUC-Rio, Depto.Engenharia Elétrica, 2007.

[2] BICK, ECKHARD, **The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework**, Aarhus: Aarhus University Press, 2002 - p. 188.

[3] GAMMA, E., HELM, R., JOHNSON, R., VLISSIDES, J., **Padrões de Projeto**, Editora Bookman, 1994.

[4] KILGARRIFF, A., GRENFENSTETTE, G., **Introduction to the Special Issue on Web as Corpus**, Computational Linguistics n#29, 2003

[5] KILGARRIFF, A., RYCHLY, P., SMRZ, P., TUGWELL, D., The Sketch Engine, **Prodeedings from the Euralex 2004**, França, p. 105-116. – 2004.

[6] KRENN, B., EVERT, S., **Can we do Better than Frequency? A Case Study on extracting PP-Verb collocations**, Proceedings of the ACL Workshop on Collocations – 2001 – Toulouse, França

[7] MANNING, C. D., SCHUTZE, H. **Foundations of Statistical Natural Language Processing**, The MIT Press, Cambridge, Massachusetts, Londres – Inglaterra, 1999.

[8] OLIVEIRA, C. FREITAS, M.C., GARRÃO, M., NOGUEIRA, C., ARANHA, C.; **A Extração de Expressões Multivocabulares: Uma Abordagem Estatística**; Revista paLavra nº12, Departamento de Letras da PUC-Rio, 2004