DESCRIÇÃO DE PALAVRAS COMPOSTAS PARA PROCESSAMENTO AUTOMÁTICO DA LINGUAGEM NATURAL

Tatiani Ramos (UFES)
tateletras@yahoo.com.br

INTRODUÇÃO

O léxico categoriza as coisas sobre as quais queremos nos comunicar, fornecendo unidades de designação, as palavras, que utilizamos na construção de enunciados. Este léxico que constrói enunciados, forma palavras, nas quais se apresentam em estruturas simples e compostas.

Uma palavra simples é uma seqüência construída sobre o alfabeto e uma palavra composta é uma seqüência de palavras simples. Com esta definição podemos distingui uma palavra simples *cara* de uma composta *cara de pau.*

Em nossa língua encontramos muitas variações lingüísticas, e isto dificulta a compreensão de algumas estruturas.

Mattoso Câmara (1987, p. 17) afirma que: um dos percalços mais sérios com que se tem defrontado os estudos lingüísticos, é o fato da enorme variabilidade da língua no seu uso num momento dado.

Ela varia nos espaços, criando no seu território o conceito dos dialetos regionais.

A existência de vários dialetos regionais faz com que somente os falantes nativos de uma determinada língua possam depreender o sentido da estrutura, porque contam com o fator de ordem pragmática. Desta forma cria-se um grande *déficit* na integridade lingüística e assim compreendemos que não basta entender e internalizar a estrutura do léxico, há também necessidade de conhecer a língua em uso.

Sendo assim, observou-se uma grande necessidade de um estudo mais detalhado sobre as expressões criadas por esta variedade, para que se consiga igualar as diferenças e facilitar a interpretação destas estruturas por todos.

NOÇÃO DE COMPOSICIONALIDADE

Neste trabalho a estrutura eleita foi a de palavras compostas, pois é a que apresenta maior grau de dificuldade de compreensão, quando observadas sobre os aspectos sintáticos, morfológicos, semânticos e pragmáticos. No que diz respeito à composição de palavras, Smarsaro (2004, p. 78) afirma que:

> A noção de composicionalidade tem a ver com a possibilidade de deduzir o significado de uma seqüência a partir dos significados dos componentes. Deduzir quer dizer calcular por um processo que pode ser formalizado.

A noção de composicionalidade apresentada acima, ainda não é tão homogênea entre os lingüistas, pois além de não darem muita ênfase para o problema da composicionalidade, cada um adota um conceito diferente gerando desta forma um estudo superficial que atende a problemática deste item lexical.

A necessidade de se chegar a um consenso sobre esta estrutura encontra-se no processo de adequação para o uso em dicionário eletrônico. Ranchhod (2001, p. 14) especifica que:

> O dicionário eletrônico é um léxico computacional concebido para ser usado, sem intervenção humana, por programas informáticos em diversas operações de processamento de linguagem natural: reconhecimento de unidades lexicais simples e complexas num texto a ser automaticamente indexada, análise de um texto para extrair informação ou para o traduzir para outra língua.

Toda esta preocupação em adequar estruturas para que possam compor uma base de dados em dicionário eletrônico, encontra-se na crescente utilização de computadores para produzir os tipos mais variados de textos.

Com o crescimento dos avanços tecnológicos as maiorias dos textos passaram a ser produzido por computador, hoje em todos os lugares nos deparamos com este tipo de serviço (supermercados, bancos, lojas, textos escolares, dissertações etc.) e isto nos coloca em contato com o uso da língua propriamente dita.

Baseado nos aumentos de textos sendo elaborados pela máquina tornou-se necessário voltarmos as nossas atenções para as expressões idiomáticas, pois apesar de muitas delas não fazerem parte

da língua padrão está sendo cada vez mais utilizadas pelos meios de comunicação para facilitar a interação com o leitor.

Um exemplo de interação com o leitor foi o *outdoor* da Grafitusa em 12/2005 na cidade de Vitória (ES), que exibiu a seguinte mensagem: "Encerramos o ano com *chave de ouro:* de *cara nova*". Nesta frase foram usadas duas expressões *chave de ouro e cara nova,* esta propaganda vem corroborar em nossa argumentação no sentido de que as expressões idiomáticas não pertencem a um escopo marginal da natureza lingüística.

Diante deste avanço na produção textual tornou-se imprescindível uma conexão entre o meio lingüístico e o computacional. Em que os lingüistas trabalham na descrição e formalização de palavras transformando-as em base de dados e os técnicos computacionais inserem os mesmos no sistema, para formar um dicionário eletrônico.

Assim se confirmou a grande importância da lingüística computacional que é a área de conhecimento, que explora as relações entre lingüística e informática, tornando possível à construção de sistemas com capacidade de reconhecer e produzir informação apresentada em linguagem natural. O tratamento automático das línguas naturais obriga a uma descrição sistemática e completa das línguas.

Esta descrição é feita de modo a considerar os processos morfológicos, sintáticos e semânticos de estruturas utilizadas por falantes nativos na construção de um dicionário eletrônico. Ranchhod (1999, p. 14) explicita de forma clara a importância desse tipo de dicionário:

> A finalidade dos dicionários eletrônicos faz com que eles tenham de ser fundamentalmente diferentes daqueles que são elaborados para utilizadores humanos, mesmo quando estes se encontram em um suporte magnético ou óptico a fim de poderem ser consultados em ambiente informatizado. Elaborados com o objetivo específico de serem usados em análise automática de texto, estes dicionários têm de conter informações lingüísticas codificadas e formatadas, pois só assim se tornam, acessíveis aos programas de análise lexical e sintática. Não podem conter lacunas (nem lexicais, nem descritivas) e todas as informações lingüísticas têm de estar coerentemente estruturadas.

O que difere o dicionário eletrônico das versões digitais está no fato de que o segundo é exatamente idêntico às tradicionais edições em papel desses mesmos dicionários que são consultados por

humanos e não podem em caso algum ser diretamente explorado por programas de análise automática de textos.

A proposta de Processamento da Linguagem Natural é possibilitar, com os recursos de que dispõe, utilizar ao máximo o que é previsível e determinado dentro da língua e explorar o que ela oferece em termos de interpretação e expressividade, nem sempre previsíveis, pois dependem também de fatores extralingüísticos, como a situação e o conhecimento compartilhado.

A IMPORTÂNCIA DO LÉXICO

Com isso o léxico surge com um fator de crucial importância em qualquer sistema de processamento automático de texto, pois é através dele que se dá a formação de palavras simples e complexas que são utilizadas na descrição.

As unidades textuais de mais baixo nível são as palavras. Assim, a primeira fase de tratamento de um texto passa inevitavelmente pela sua análise lexical. E isto consiste em:

- Identificar as unidades lexicais do texto;
- Descrever cada uma delas através de informações lingüísticas formalizadas, à cabeça das quais se encontram as de natureza morfológica e categorial;
- Resolver o maior número de ambigüidades lexicais provocadas pela homografia.

Separar as unidades lexicais de palavras simples como andar, modelo, par e passo não é complicado: elas são separadas tipograficamente um das outras. Porém, muitas unidades lexicais não são palavras simples. São palavras compostas como a estrutura *par e passo*, formada por duas palavras simples.

O sistema de etiquetagem das palavras do texto, isto é, a associação de informações lingüísticas, depende crucialmente da noção de palavra simples e composta. Se o sistema de etiquetagem não é utilizar a noção de composto a palavra par, por exemplo, receberá o mesmo tratamento quer em:

(1) Cada um dançou com o seu par.

em que par é uma palavra simples e pluralizável, ao contrário do outro exemplo em:

(2) Todas as decisões foram tomadas *a par e passo*.

em que o mesmo elemento perde as propriedades (sintáticas e morfologias) que tem em (1). Ele só tem valor dentro da combinação *a par e passo*, que no seu todo se comporta como um advérbio e deve ser etiquetada como tal.

Como nos exemplos acima existem inúmeros casos de expressões compostas com estruturas de: Nome e Nome, Nome adjetivo, Nome e sintagma verbal, etc. e para os identificar de forma fundamentada é necessária utilizar alguns critérios lingüísticos como afirma Ranchhod (1990, p. 49) e Baptista (1995, p. 66):

> ...que vão desde a análise do seu comportamento morfológico, até à verificação da sua, total ou parcial, perda de composicionalidade, lexical, sintática e semântica. Esses critérios são igualmente necessários para distinguir os nomes compostos de grupos nominais livres.

Uma seqüência será considerada composta, se apresentar restrições quanto às propriedades sintáticas em relação à outra seqüência com a mesma categoria de palavras.

É na intersecção dos vários critérios que se define a composicionalidade de uma dada combinação que será tanto mais fixa, quanto mais restrições se observarem em relação às propriedades sintáticas que caracterizam um grupo nominal livre formado pela mesma seqüência interna de categorias gramaticais.

Estes procedimentos são essenciais para a codificação de dados na formação de um dicionário eletrônico, pois o processamento automático necessita, antes de mais, de um tratamento das unidades básicas da língua (palavras simples e compostas) diferente do que é feito pelos dicionários de uso informatizados.

As expressões cristalizadas ou fixas, segundo Vale (2001, p. 130),

> São expressões formadas por mais de um segmento (um segmento compreendido, em língua escrita, como uma seqüência de letra delimitada por dois separadores, qual seja: um espaço em branco, um sinal de

pontuação) cujo significado total não se pode ser deduzido pelo significado das partes que a compõem.

Por exemplo, a expressão "*fazer uma vaquinha*" não significa que alguém irá confeccionar uma vaca, mas sim juntar uma certa quantidade de dinheiro.

Este tipo de expressão é muito comum na nossa língua e aparece em várias estruturas como as substantivais, ex: "*plano de saúde*", as adverbiais, ex: "*dos pés a cabeça*", com verbo-suporte "*fazer gato e sapato*", etc. De uma maneira geral as expressões cristalizadas (EC) se estruturam como frases comuns, mas que precisam de um contexto que impeça as interpretações de forma errônea, ou seja, de modo partitivo.

O dicionário eletrônico colabora para sanar com problemas gerados pela semântica das palavras ou estruturas, o que dificulta o manuseio do idioma por estrangeiros que buscam ter o contato com textos escritos, e isto funciona como uma forma de igualar as diferenças tornando as leituras e interpretações textuais mais completas nas línguas em que estão sendo elaboradas pesquisas para a implementação do dicionário eletrônico.

Na constituição de um dicionário eletrônico, a listagem das formas deve ser a mais completa possível. Isto diz respeito tanto às formas canônicas como as formas flexionadas. Para uma língua como o português, um dicionário destinado à utilização em máquina deve explorar de modo aprofundado as estruturas flexionadas e isto se refere à flexão verbal e também a outras classes que tenham essas propriedades como os substantivos e os adjetivos.

É importante ressaltar que se faz necessário observar as propriedades morfológicas, sintáticas e semânticas das estruturas que obedecem aos critérios de aceitação e não-aceitação, realizados por pesquisadores franceses de lingüística (Laporte, 2000) e brasileiros (Vale, 2001; Smarsaro, 2004) os estes testes são realizados no sentido de cercar os limites entre a aceitação ou não de outros itens lexicais alocados no interior da estrutura, para verificar a sua fixidez.

Estes testes são utilizados na análise das ocorrências, levando em conta as distribuições sintáticas dos componentes de cada seqüência e a interpretação lingüística dos falantes nativos. Os critérios

utilizados que se referem às propriedades sintático-semânticas e morfológicas são os seguintes:

- Variação em gênero;
- Variação em número;
- Variação em grau;
- Substituição de parte do SN por outro;
- Substituição do determinante do SN;
- Redução ou elipse de SN;
- Inserção de um item lexical no grupo nominal;
- Coordenação do adjetivo com outro adjetivo;
- Inserção de um advérbio na EC.

Com base nestes critérios seguem-se aqui uns exemplos que representam os três primeiros pontos como forma de demonstração dos critérios

Variação em Gênero

No que se trata da variação em gênero, pode ocorrer ou não um bloqueio da EC.

a) – João está com uma *cara porca.*

b) – Ana está com uma *cara porca.*

c) * – João está com *um cara porco.*

d) * - Ana está com *um cara porco.*

A mudança de gênero no exemplo 1, efetuada no sujeito masculino para o feminino não apresenta bloqueio na EC. Para os exemplos 1-(c) uma simples variação do artigo interferiu no grau de fixidez da EC, pois em 1-(b) o sentido aponta para um rosto com a expressão de safado e em 1- (c) entende-se que João está acompanhando um homem sujo.

Variação em Número

2 – João está com uma *cara porca.*

a – Ana e João estão com *uma cara porca.*

b - * Ana e João estão com *umas cara porca.*

c - ? Ana e João estão com *umas caras porca.*

d – Ana e João estão com *umas caras porcas.*

A expressão ao ser testada com o critério 2, demonstra cumplicidade com o artigo, pois na frase 2.a o artigo está no singular e a expressão permanece no singular.Já para a frase 2.d o artigo foi para o plural e a expressão também, o que torna aceitável as frases 2b e 2c.

Variação em Grau

No ponto em que se encontra nossa pesquisa os testes em grau só foram efetuados no modo sintético com o sufixo -inho.

João está com uma *cara porca.*

a- * João está com uma *cara porquinha.*

b- * João está com uma *carinha porquinha.*

c- João está com uma *carinha porca.*

O sufixo -inho indica afetividade, quando acrescentado a uma palavra, mas para as frases acima somente a 3.b tornou-se aceitável, a 3.c torna-se duvidosa, pois para o falante nativo é aceitável, mas lingüisticamente ela foge as regras. Para a frase 3.a é totalmente inaceitável a variação na última palavra em se tratando de uma palavra composta.

Observamos também a variação em grau dessas expressões no modo analítico.

João está com uma *cara porca.*

*João está com uma *pequena* cara porca.

*João está com uma *minúscula* cara porca.

*João está com uma *grande* cara porca.

*João está com uma *enorme* cara porca.

As formas "pequena cara porca" e "grande cara porca" causam estranhamento não por questões de lingüística, mas sim por questionar o tamanho. Para se dizer que alguém tem a *cara porca* não precisa indicar o tamanho.

Os exemplos demonstrados acima geram a formalização das estruturas para a inserção em dicionário eletrônico. Segue-se abaixo uma pequena amostra dos processos de formalização:

FORMALIZAÇÃO DAS ESTRUTURAS

Procuramos apresentar uma formalização das classes Nome adjetivo (Nadj.), Nome preposição e nome (N de N), descritas a partir de critérios morfossintáticos e semânticos no 1º semestre.

Classe	Exemplos	Número de Ocorrências
Nadj	Cara porca	31
N de N	Cara de sono	49

A formalização consiste em apresentar uma descrição codificada das propriedades estruturais (morfológicas, sintáticas e semânticas) para inserção em programas de dicionário eletrônico.

Para a classe Nome adjetivo (Nadj.) a simbologia utilizada é resultado da aplicação de critérios.

"N" é usado para indicar a categoria de substantivos.

Quando a seqüência (N) for adjetiva, anota-se "Nadj.", quando não é, anota-se "Nsubst." Ex: Cara porca (Nadj.) ou cara honesto (Nsubst.). Junto a esta seqüência ocorrerá ou não a função predicativa, quando sim , anota-se "+f.pred.", quando não, anota-se "-f. pred.. Ex: Cara feia (Nadj. + f. pred.) ou cara larga (Nadj. – f.pred.).

Quando o resultado da aplicação de critérios apresentar variação em gênero, anota-se "+g"; quando não apresenta, anota-se "-g". Os códigos (+g ou –g) indicam que os substantivos podem ser flexionados em gênero. Ex: Cara amarrada (Nadj. + f. pred. –g).

Quando a seqüência (Nadj.) não apresenta a variação em número, anota-se "-n", quando apresenta, anota-se "+n", sendo que este "n" se dividirá em "n^1" e "n^2", indicando a variação por partes dentro da própria seqüência. Ex: Cara suja (Nadj. – f. pred. – g + n) ou cara legal (Ndj. + f.pred. – g + n^1 + n^2).

Quando a seqüência (Nadj.) não apresentar variação em grau, anota-se "- Ngr", quando sim, anota-se "+Ngr" e nesta simbologia acrescenta-se ainda os números 1 e 2 que indicam a parte que irá variar em grau. Ex: Cara azeda (Nadj. + f. pred. – g + n^1 + N^1gr) ou cara triste (Nadj. + f.pred. – g + n^1 + n^2+ N^1gr + N^2gr

A aplicação de critérios também levou em consideração se "n^1" tem "n^2", quando sim, anota-se "+ter"; quando não, anota-se "-ter". Os códigos (- ter ou +ter) indicam a representação da existência de uma frase simples N^2 ter N^1, na qual N^2 e N^1 conservam o mesmo sentido da palavra composta N^1 de N^2.

Ex.: Cara lavada (Nadj. + f. pred. – g + n^1 + n^2 + N^1gr – ter + n: fs).

CONSIDERAÇÕES FINAIS

A realização de um dicionário eletrônico com estas propriedades se configura como uma condição para a melhoria da qualidade de programas computacionais, que lidam com processamentos da linguagem natural.

Isto se faz necessário, pois em várias línguas existem muitas palavras com características de composição e a descrição de nomes compostos pode, possivelmente, solucionar grande parte de um dos problemas clássicos no processamento das línguas-reconhecimento de formas numa seqüência linear sem o comprometimento do sentido das informações, eliminando-se ambigüidades, redundâncias, repetições e informações agramaticais.

Sendo assim é possível organizar uma metodologia para o português que abrange os para falantes nativos e para estrangeiros fundamentada em melhores condições didáticas.

REFERÊNCIAS

GARRÃO, Milena de Uzeda. *Um estudo de expressões cristalizadas e sua inclusão em um tradutor automático bilíngüe* (*português/inglês*)*: o caso de "bater + SN"*. 2001. Dissertação de Mestrado. Rio de Janeiro: PUC-RJ, 2001.

SMARSARO, Aucione das Dores. *Descrição e formalização de palavras compostas do português do Brasil para elaboração de um dicionário eletrônico*. Tese de doutoramento. 2004. Rio de Janeiro: PUC-RJ, 2004.

VALE, Oto Araújo. *Expressões cristalizadas do português do Brasil: uma proposta de tipologia*. Tese de doutoramento. 2001. Araraquara: UNESP, 2001.

CAMARA, Joaquim Mattoso. *Estrutura da língua portuguesa.* 17ª ed. Petrópolis: Vozes, 1987.

BASÍLIO, Margarida. *Formação de palavras no português do Brasil.* 1ª ed. São Paulo: Contexto, 2004.

RANCHHOD, Elizabete Marques. *Tratamento das línguas por computador: Uma introdução à lingüística computacional e suas aplicações.* 1ª ed. Lisboa: Tipografia Louzanense, 2001.

BAPTISTA, J.M.E. *Estabelecimentos e formalização de classes de nomes compostos.* Dissertação de Mestrado. Lisboa. 1994.