A HOMONÍMIA E O COMPUTADOR

Claudia ZAVAGLIA (Universidade Estadual Paulista).

ABSTRACT: *This paper is an attempt to describe and to discuss the behaviour of the phenomenon of the homonymy in tests carried out in grammar checkers, as well as the possibility of desambiguation for some linguistic forms.*

KEY WORDS: *Homonymy, desambiguation, computational lexicon*.

Gardner (1995:197,198) chamou a atenção para certas intuições que os falantes de uma língua possuem quando afirmou:

> "O fato de termos intuições claras sobre idéias aparentemente sem sentido como 'Idéias verdes incolores dormem furiosamente' serve de base às principais contribuições de Chomsky à lingüística e, no que se refere a isto, à ciência cognitiva em um sentido mais amplo. O que Chomsky conseguiu na sua monografia, e nas inúmeras obras que se seguiram, foi chamar a atenção para certas propriedades das sentenças que os falantes e ouvintes normais conhecem intuitivamente, mas que derivam de uma compreensão mais profunda da língua cujas propriedades podem ser explicitamente conhecidas apenas pelos lingüistas. (...) Ele indicou e sugeriu mecanismos subjacentes à habilidade humana para detectar e resolver ambigüidades em senteças como *Flying planes can be dangerous* [Pilotar aviões pode ser perigoso ou Aviões voando podem ser perigosos] (...)".

Partindo-se de tais considerações, poderíamos dizer que o falante estaria apto a decodificar e codificar certas ambigüidades de uma língua natural geradas pela polissemia e a homonímia inconscientemente.

Uma das maiores ambições de lexicólogos, lexicógrafos e lingüistas computacionais é a tentativa de se demarcar as fronteiras entre a polissemia e a homonímia, ou seja, quando um item lexical é polissêmico e quando, ao contrário, é homônimo. A delimitação do campo de ação da significação de uma unidade lexical torna-se necessária para que se possa distinguir as unidades hmônimas das polissêmicas.

Segundo Biderman (1991):

> "A polissemia é um fenômeno que ocorre no interior das redes de significação do léxico geral da língua comum, em virtude da economia lingüística, com o reaproveitamento freqüente de um certo número de lexemas no processo de comunicação. A ampliação do uso de uma palavra e a metaforização contínua da linguagem acarretam a freqüência de muitas unidades lexicais, gerando a polissemia. Inversamente, são palavras homônimas as unidades lexicais em que significantes morfofonicamente idênticos têm significados diferentes".

O critério mais utilizado para se fazer a distinção de um item lexical homônimo de um polissêmico é a verificação da origem etimológica do mesmo, ou seja, um critério diacrônico. Considerando-se que, em português, as pesquisas etimológicas de itens lexicais são escassas e insuficientes para que possam oferecer segurança e credibilidade, no que diz respeito à origem de uma palavra, torna-se difícil adotar como critério básico de identificação de um item lexical, o estudo diacrônico do mesmo.

O dicionário "Aurélio", o mais conhecido e utilizado por falantes da língua portuguesa, adotou, como recurso lexicográfico, para distinguir formas homônimas das polissêmicas, a etimologia das mesmas. Assim, somente quando a diacronia pôde oferecer étimos diferentes para uma mesma forma lexical, a mesma foi considerada homônima. Tal procedimento lexicográfico fez com que palavras tais como *banco*$_1$ e *banco*$_2$, *ponto*$_1$, *ponto*$_2$, *ponto*$_3$ e *ponto*$_4$ (Cf. Biderman, 1991) fossem consideradas polissêmicas quando, na verdade, são homônimas.

Dessa forma, não sendo suficiente o critério diacrônico para estabelecer a oposição entre homonímia e polissemia, buscar-se-á tal diferenciação com base na semântica. Desse modo, entende-se por homônimos aqueles significantes que não possuem nenhum sema em comum e por polissêmicos aqueles que possuem ao menos um sema em comum, por exemplo, *'manga'* (parte do vestuário) / *'manga'* (fruto) em que não temos nenhum traço semântico unindo os dois significantes e logo, são homônimos e *'boca'* (cavidade na parte inferior da cabeça) / *'boca'* (abertura de garrafa) em que temos o traço 'espaço oco' pertencente aos dois significantes e, portanto, polissêmicos.

No português, a homonímia pode ocorrer entre categorias gramaticais idênticas: 'manga'(substantivo), entre categorias gramaticais distintas: 'abandonado' (substantivo e adjetivo), 'visto'(substantivo, particípio passado e preposição), 'canto' (substantivo e verbo) entre outras.

O choque homonímico entre nome/adjetivo, nome/verbo e adjetivo/particípio passado verbal possui uma alta freqüência na língua portuguesa[1]. É interessante notar, porém, que nos livros em que se encontram definições/estudos/pesquisas sobre a homonímia, os exemplos de homônimos citados pelos autores são sempre os mesmos, a saber: 'são', 'manga', 'canto', 'cabo', 'alimento', entre "poucos" outros. Assim sendo, um levantamento de

formas homônimas existentes no português do Brasil faz-se cada vez mais necessário, seja para atender lexicólogos e lexicógrafos seja para lingüistas computacionais[2].

O fenômeno da homonímia das línguas naturais tem sido motivo de vários empecilhos para o desenvolvimento do Processamento Automático das Línguas Naturais - PALN - em Lingüística Computacional para o qual um dos grandes desafios é tentar transportar para a máquina as suas delimitações, uma vez que a mesma não possui "intuições" interpretativas como os humanos.

O presente trabalho objetiva demonstrar tal dificuldade, através de exemplos homonímicos conflitantes, em análises de textos escritos realizadas por revisores gramaticais existentes para a língua portuguesa do Brasil, no caso, aquele elaborado pelo convênio USP-São Carlos e Itautec-Philco S/A[3].

O revisor gramatical tem por meta, através de uma análise sintática automática, identificar desvios lingüísticos que fujam ao padrão da norma culta da língua portuguesa do Brasil em textos escritos. O repertório lexical, que é a base lingüística de tal revisor, contém diversos homônimos classificados a partir de um critério morfológico que contempla as suas várias classificações em uma mesma entrada ou em entradas diferentes, tal como a forma "fala", categorizada como substantivo feminino singular, presente indicativo, 3ª pessoa do singular e imperativo afirmativo, 2ª pessoa do singular.

A partir de um *corpus* de aproximadamente 10.000 ocorrências entre textos literários, jornalísticos e técnicos, foram realizados testes com o *software* visando a qualificar a performance do mesmo e o tratamento dado às formas homônimas.

A ferramenta computacional assim funciona: detectada uma concordância indevida, ela "seleciona" a frase e interage com o usuário sugerindo-lhe uma recomendação na qual demonstra qual o tipo de problema "detectado". O usuário pode, se sentir necessidade, buscar mais informações gramaticais (denominado, no arquivo, de 'Mais informações') sobre tal desvio lingüístico. Por exemplo, em um dos textos que a máquina analisou foi selecionada a seguinte frase: *Seus dois fundos de renda fixa foram os que mais renderam no ano*, em que o *software* oferecia a seguinte recomendação: "Se 'fixa' estiver se referindo a 'fundos', verifique a concordância de número". Em tal recomendação não se compreende exatamente a qual tipo de concordância de número se refere a máquina e para tanto, buscamos o 'Mais informações' do revisor que nos ofereceu o seguinte informe: "Em português, o verbo deve concordar em número e pessoa com o sujeito da sentença". A partir de tais informações entende-se que o revisor considerou a lexia 'fixa' como sendo uma forma verbal, mais especificamente aquela da terceira pessoa do singular do presente do indicativo e desse modo, sugeriu a concordância verbal entre 'fundos' e 'fixa', desconsiderando, portanto, a categoria adjetivo da mesma e a forma verbal 'foram' que a seguia. Tal procedimento é chamado de 'falso erro', ou seja, uma intervenção indevida realizada pelo revisor, já que a frase não possui nenhum erro. De fato, trata-se de uma ambigüidade interpretativa da ferramenta gerada, neste caso, pela homonímia categorial, uma vez que 'fixa'

encontra-se categorizada como forma verbal do presente do indicativo e imperativo afirmativo do verbo ‘fixar’ na base lingüística do mesmo. O procedimento lingüístico-computacional utilizado para descaracterizar e desfazer, conseqüentemente, tal ‘falso erro’ foi o de realizar a concordância de número tanto com a lexia ‘fundos’ quanto com a ‘renda’ no interior de tal sintagma nominal; havendo concordância de número dentro dos padrões lingüísticos da língua portuguesa com uma das duas formas o revisor foi induzido a não mais acusar nenhum tipo de desvio. Percebe-se que tal ‘problema’ foi sanado pontualmente, ou seja, especificamente para este caso, mas será que estão eliminadas interferências deste tipo por parte do corretor, satisfatoriamente, para outros contextos? Além de existir um choque homonímico em tal frase, trata-se também da presença de uma lexia complexa, a saber: ‘fundos de renda fixa’, a qual poderia ter sido levada em consideração, pelo revisor, como uma entrada independente e portanto, categorizada como substantivo masculino plural, o que não teria conduzido a máquina ao ‘falso erro’.

Um outro exemplo evidencia o choque homonímico *substantivo X adjetivo* na seguinte frase: *Mataram o cara e deu esse problema*, em que o revisor selecionou o sintagma nominal “o cara” e recomendou: “se ‘o’ estiver se referindo a ‘cara’, verifique a concordância de gênero” e o ‘Mais informações’ instruiu: “O artigo, o pronome, o numeral e o adjetivo determinantes devem sempre concordar em gênero e número com o substantivo a que se referem”. No repertório lexical do revisor, implementou-se o lexema ‘cara’ somente como forma feminina do adjetivo ‘caro’, uma vez que na implementação automática das informações gramaticais aos itens lexicais, visando uma categorização que fosse a menos complexa possível para se tentar evitar a ambigüidade gramatical, alguns critérios classificatórios para a homonímia categorial foram criados objetivando atenuar os problemas das regras gramaticais do *software*. Desse modo, certas classes gramaticais foram unidas em uma única categoria e assim, para qualquer substantivo que fosse também adjetivo, optou-se pela segunda categoria respectivamente, já que a maioria dos adjetivos pode ser substantivada. O problema tornou-se mais grave ainda, devido ao duplo gênero que pode ter a lexia ‘cara’ enquanto substantivo, qual seja: feminino, significando “rosto, semblante” e masculino, denotando “indivíduo”. Na substantivação do adjetivo ‘cara’ a máquina não é capaz de atribuir-lhe o artigo masculino, uma vez que ela obtém a informação de que tal forma é feminina, e não lhe resta outra alternativa a não ser acusar a concordância errônea. A desambigüização é alcançada através do acréscimo da categoria substantivo masculino, na base lingüística do revisor, para a lexia ‘cara’. O fato de ter-se determinado a escolha da categoria adjetivo, na implementação computacional, de todas as formas homômimas *substantivo X adjetivo* demonstra a insuficiência de tal critério. Se, por outro lado, formas homônimas tivessem sido implementadas a partir de um critério lexicoestatístico, tais problemas estariam fadados a diminuirem, ou até mesmo a desaparecerem. Biderman (1996) tece comentários sobre a pesquisa realizada

pela Universidade de Lisboa, para a elaboração de um vocabulário fundamental do português, em que analisa a metodologia e os critérios adotados pelos pesquisadores para tal empreendimento e nos declara: "A análise dos dados revelou resultados interessantes. Em casos de homonímia *substantivo X adjetivo - amigo (substantivo) X amigo (adjetivo)*; *jovem (substantivo) X jovem (adjetivo)*; *ideal (substantivo) X idela (adjetivo)* -, a apreciação empírica dos dados induziria à categorização como adjetivo, a categoria primeira. Ora, a análise dos contextos dessas e outras formas homógrafas revelou que os substantivos são mais freqüentes". Dessa forma, acreditamos que buscas de palavras homógrafas e da freqüência de ocorrência das mesmas seja um caminho para a resolução de alguns tipos de ambigüidade.

Pondo em relevo a face quantitativa da linguagem, Biderman (1978) ressalta que "a freqüência seria uma característica tão típica do signo como os traços distintivos que o opõem aos demais elementos do sistema". Endosso suas palavras, reiterando mais uma vez o dito anteriormente em trabalho recente (Cf. Biderman, 1996:28):

> "Dada a enorme extensão do léxico, uma seleção lexical criteriosa e baseada em princípios lexicoestatísticos apresentou-se como a melhor alternativa para estabelecer os *indices verborum* das palavras mais freqüentes e usuais dentre as centenas de milhares que constituem o léxico de uma língua de civilização moderna. Dessa forma, podem-se evitar o empirismo e uma seleção vocabular com base apenas na intuição".

Nossos testes enfatizam, ainda, a veracidade de tais afirmações na seguinte frase em que a máquina detectou problemas: *"Tivemos de importar acrílicos sem emenda, tintas especiais e usar materiais sintéticos, como o córean, que tem a aparência de pedra, mas é perfeitamente moldável", explica o empresário Marcos Brochini, um dos sócios da indústria de móveis Positano, que desenvolve a coleção* em que evidencia um problema com a lexia 'tinta'; na frase, um substantivo e no repertório lexical, categorizada como adjetivo. A verificação de tal lexia no Dicionário de Freqüências do Português Contemporâneo (DFPC) nos revela: 120 ocorrências do lema 'tinta' são categorizadas como substantivo, sendo 68 para a forma do singular e 52 para a forma do plural; a forma adjetiva de 'tinta' não ocorreu nenhuma uma vez. Claro está, portanto, que a lexia 'tinta' é mais freqüente como substantivo e assim deverá ser implemetada no *software* para a sua desambigüização. Vejam-se as frases: "*Candidato a uma vaga de gerente financeiro, não ficou mais do que duas semanas sem emprego"* / "*Mesmo assim, o tema causa polêmica"* / "A *ventilação injeta oxigênio no pulmão por meio do controle da pressão, fluxo e quantidade do gás, essencial para o funcionamento das células do organismo" / "Todas as regiões registraram queda em julho"*. Temos aqui problemas detectados pelo revisor que envolvem as lexias

homônimas *substantivo X adjetivo*, a saber: ‘vaga’, ‘polêmica’, ‘fluxo’ e ‘queda’ todas categorizadas como adjetivo na base lingüística da máquina e empregadas como substantivo nas frases. Para tais formas, deverá ser realizada uma busca em *corpora* representativos da língua portuguesa (variante brasileira) a fim de se verificar qual categoria é a mais freqüente, já que no DFPC encontramos uma alta freqüência de ocorrência para as mesmas vinculada ao choque homonímico, ou seja, não podemos precisar qual é a categoria individual mais freqüente sem analisá-las contextualmente, a não ser para a lexia ‘queda’. A mesma é mais freqüente como substantivo do que como a forma feminina do adjetivo ‘quedo’, como constatamos: de 315 ocorrências da lexia ‘queda’, 295 são como substantivos na forma singular e as outras se dividem entre a forma plural e lexias complexas. Com a introdução da categoria substantivo, no repertório lexical, para tais lexias, ocorrerá a desambigüização das mesmas.

Em casos de homonímia *substantivo X verbo*, a freqüência das formas também indicará com precisão qual é a categoria mais freqüente. Na frase *“Vou visitar amigos e inimigos para pedir ajuda”* o revisor detectou um ‘falso erro’ devido à lexia ‘ajuda’, que se encontra classificada no repertório lexical como forma verbal. No DFPC ‘ajuda’ consta de 569 ocorrências como substantivo, indicando-nos claramente a sua alta freqüência como tal categoria. Sendo que o DFPC foi elaborado com base em um *corpus* de 6 milhões de ocorrências de palavras, a freqüência das lexias aqui analisadas é considerada alta e representativa para os índices numéricos demonstrados anteriormente.

As considerações e os resultados aqui discutidos nos levam a inferir o quão necessária é a pesquisa sobre o fenômeno da homonímia de uma língua natural. Em se tratando de implementações computacionais, acreditamos que o levantamento estatístico poderá resolver algumas questões sobre a ambigüidade da linguagem natural, já que serão identificadas as combinatórias gramaticais de formas ambíguas em *corpora* representativos da língua portuguesa (UNESP-Araraquara). Dessa forma, presume-se que poderão ser identificados os contextos lingüísticos nos quais se encontram as formas homógrafas mais freqüentes, para que então se possa oferecer subsídios lingüísticos senão para a resolução, pelo menos para o abrandamento, da ambigüidade lingüística.

NOTAS

[1]Levantamentos de tal foram realizados no Dicionário de Freqüências do Português Contemporâneo de Maria Tereza Camargo Biderman, UNESP, Araraquara. [2]Pesquisa de doutorado de Claudia Zavaglia, UNESP, Araraquara, sob orientação de Maria Tereza Camargo Biderman. [3]Projeto realizado no NILC - Núcleo Interinstitucional de Língüística Computacional - USP, São Carlos, SP.

RESUMO: *Este texto objetiva descrever e discutir o comportamento do fenômeno da homonímia em testes realizados em revisores computacionais e a possibilidade de desambigüização para certas formas.*

PALAVRAS-CHAVE: *Homonímia, desambigüização, léxico computacional*.

REFERÊNCIAS BIBLIOGRÁFICAS:

BIDERMAN, M.T.C.(1991). Polissemia Versus Homonímia. *Anais do XXXVIII Seminário do GEL*. Franca. ----- (1978). *Teoria Língüística: lingüística quantitativa e computacional*. Rio de Janeiro: Livros Técnicos e Científicos. ----- (1996) Léxico e Vocabulário Fundamental. *Alfa*, **n.40**. São Paulo, GARDNER, H. (1995). *A Nova Ciência da Mente*. São Paulo: EDUSP.