2
O Léxico e o Processamento de Linguagem Natural

2.1. O desempenho dos sistemas para o processamento das línguas

Atualmente, muitos usuários de computadores estão familiarizados com vários produtos comerciais, cuja função é processar textos escritos: editores de textos, sistemas de busca de páginas na *web*, sistemas de ajuda à tradução etc. Segundo Laporte (2000), esses exemplos de sistemas computacionais estão disponíveis e é fácil constatar que, embora sejam úteis, o seu desempenho ainda não é satisfatório. Os melhores editores de textos apontam erros em palavras corretas, propõem correções erradas e deixam de detectar certos tipos de erros ortográficos. Os sistemas de busca na *web* selecionam, às vezes, dezenas de páginas sem qualquer relação com o assunto pesquisado pelo usuário, mesmo que este expresse seu objetivo de forma suficientemente precisa. Até os textos produzidos pelos melhores sistemas de ajuda à tradução necessitam de um sistema para tradutores humanos, devido a erros de tradução que, aliás, tornam os resultados da tradução automática quase um gênero literário cômico.

Retomando o que afirma M.Gross (1991:7), hoje em dia, praticamente todos os textos (livros, jornais, revistas, periódicos, mala comercial etc.) são produzidos mediante computadores. Segue-se daí, em princípio, que os arquivos podem agora ser armazenados em formato computacional. Os programas de computador podem escanear os textos desses arquivos em busca de informações específicas.

Provavelmente, esses problemas requerem mais estudos, e essas dificuldades poderão ser superadas, considerando-se que o material a ser processado é de natureza lingüística e as dificuldades podem se dar, porque há diferentes possibilidades de interpretação para as palavras: não há total regularidade, e mais: não é clara a noção de sintagma e composto e é necessário que lidemos também com as questões homonímia/polissemia. Contudo, o aumento das potencialidades técnicas dos computadores foi rápido, e enormes quantidades de textos se tornaram disponíveis em

suportes eletrônicos. Numerosos sistemas de processamento de textos foram elaborados, muitas vezes em apenas alguns meses, com aplicação de métodos e aproximações matemáticas e com pouca integração de dados lingüísticos. É por essas razões que destacamos a relevância desta pesquisa, pois a construção de um dicionário de palavras compostas, sem dúvida, poderá melhorar a qualidade dos programas que pressupõem descrições lingüísticas.

2.2
A importância do léxico para o processamento

Segundo Ranchhod (2001), nos últimos anos, tornou-se evidente que os recursos lingüísticos e, em particular os recursos lexicais, são a pedra de toque de qualquer sistema de processamento de linguagem natural. Na verdade, a crescente necessidade de aplicações da lingüística computacional fez ressaltar a carência de dados lingüísticos de dimensões reais, em particular, de léxicos e gramáticas de grande cobertura. Para atender a essas exigências de qualidade, o tratamento automático requer das línguas uma descrição sistemática e o mais completa possível, pois a insuficiência de dados lingüísticos pode gerar falhas no processamento automático. Ao falarmos de processamento de linguagem natural, surgem vários problemas, como a grande variação morfológica e sintática das unidades lexicais ou a ambigüidade intrínseca da língua portuguesa. Para resolvê-los, destacam-se três níveis de análise lingüística: morfológico, sintático e semântico. Para cada nível associam-se descrições lexicais apropriadas. Se, em relação ao tratamento do léxico, os dicionários utilizados pelos sistemas de processamento não forem adequados, quer do ponto de vista da sua cobertura lexical, quer, do ponto de vista da formalização e sistematização da informação lingüística, isso afetará não só a análise lexical de um determinado texto, mas também todas as fases de processamento subseqüentes. Se, por exemplo, uma palavra não for reconhecida ou não for corretamente identificada, a análise sintática da frase ou da estrutura em que ela se encontra não poderá ser feita. O léxico surge, portanto, como um componente de grande importância em qualquer sistema de processamento automático da linguagem natural.

2.3
Contribuições da lingüística para o processamento das línguas

Durante 60 anos, tanto a lingüística quanto a informática desenvolveram suas pesquisas de modo quase independente. Há trabalhos conjuntos, mas, pelo menos para o português do Brasil, a investigação ainda é relativamente recente. São dois mundos que ainda não se conhecem muito bem. De um lado, os lingüistas, de outro, os engenheiros. Talvez por isso o processamento da linguagem natural ainda apresente resultados insatisfatórios. As dificuldades aqui apontadas representam indicativos de que o desempenho desses sistemas disponíveis ainda é passível de melhoria e outros a serem elaborados poderão apresentar melhor qualidade no desempenho de suas funções. A observação, a descrição, a codificação das propriedades e a adequação de dados lingüísticos necessitam de um ritmo de elaboração mais lento, mas com certeza possibilitarão progressos substanciais no desempenho dos sistemas. Por outro lado, Santos (1999) argumenta que ao tentarmos resolver um dado problema (isto é, ao tentar construir um programa que manipula a língua) é que surge o momento de nos debruçarmos quer sobre algumas características do léxico ou da gramática, quer sobre as teorias que pretendem dar suporte a esse problema. Isso quer dizer que, para obtermos resultados mais eficazes no processamento automático da linguagem natural, a Lingüística e a Informática são dois campos que devem estar sempre em interação, pois a Informática necessita das descrições lingüísticas e a Lingüística, por sua vez, deve apresentá-las de modo que possam ser representadas e utilizadas pela máquina.

Em meio a essas dificuldades constatadas no processamento da língua, a elaboração de um dicionário eletrônico torna-se uma necessidade real, tanto do ponto de vista da qualidade das informações, quanto do ponto de vista da quantidade de palavras lexicalizadas na língua. Um dicionário eletrônico que apresente a descrição das palavras, no caso palavras compostas, com suas propriedades morfológicas, sintáticas e semânticas, provavelmente será um dos recursos que poderão resolver uma grande parte dos problemas de natureza lingüística encontrados no processamento das línguas.

Com esse objetivo, faz-se um recorte entre os processos de formação de palavras do português do Brasil, apresentando uma seleção de palavras candidatas a compostas, formadas por justaposição, com a estrutura NdeN, para descrevê-las e formalizá-las, tendo em vista a representação formal utilizável em dicionário eletrônico.

2.4
As propriedades de um dicionário eletrônico

Os dicionários eletrônicos fazem parte da maioria dos programas que envolvem procedimentos de reconhecimentos de unidades lingüísticas significativas M.Gross (1989). Nesse tipo de programa, um texto é submetido, inicialmente, a um procedimento de segmentação das unidades gráficas (as palavras). Em seguida, consulta-se um dicionário a fim de determinar a natureza de cada uma dessas unidades. Se uma palavra não for encontrada no dicionário, uma análise posterior mais apurada do texto será bloqueada ou embaraçada nas melhores das hipóteses. Se, em relação ao tratamento do léxico, os dicionários utilizados pelos sistemas de processamento não forem adequados, quer do ponto de vista da sua cobertura lexical, quer do ponto de vista da codificação e sistematização das sequências lingüística, isso afetará não só a análise lexical de um determinado texto, mas também todas as fases de processamento subseqüentes. Logo, todas as informações lingüísticas devem ser dadas à máquina de forma completa e explícita.

Um 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 (de natureza terminológica ou não) num texto a ser automaticamente indexado, análise de um texto para extrair informação ou para traduzir para outra língua, etc. Essa 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 suporte magnético ou óptico, a fim de poderem ser consultados em ambiente informatizado. Contudo, o fato de as versões digitais

dos dicionários de uso serem freqüentemente comercializadas com a designação de dicionários eletrônicos pode levar a uma certa confusão entre os dois tipos de léxicos, que convém esclarecer: segundo M. Gross (1989), a ambigüidade do termo informatização levou a um mal-entendido entre as duas categorias de dicionários. As informações contidas em cada dicionário não têm nada em comum: num caso, são codificações não transparentes (destinadas aos profissionais da lingüística computacional); no outro, textos destinados ao grande público. Em geral, as versões informatizadas dos dicionários de uso são completamente idênticas às tradicionais edições em papel desses mesmos dicionários: idêntico conteúdo, idêntica estruturação de entradas, idêntica cobertura lexical. A sua diferente apresentação pode facilitar a sua consulta, mas não torna diferentes os seus objetivos: em papel ou em formato digital, destinam-se a serem consultados por humanos e não podem em caso algum ser diretamente explorados por programas de análise automática de texto. Apesar de alguns aspectos comuns, há entre os léxicos computacionais e aqueles que não o são, diferenças apreciáveis.

A diferença mais evidente reside no fato de que, num dicionário de uso, as seqüências lingüística não estão codificada, enquanto que a codificação é um requisito imprescindível de um dicionário eletrônico. Os dicionários eletrônicos são aqueles elaborados com o objetivo específico de serem usados em análise automática de texto; por isso 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. As informações de natureza sintático-semântica também têm de ser tratadas nos dicionários eletrônicos, por isso esses dicionários devem ser, desde o início, concebidos para poderem receber cumulativamente não só informações adicionais sobre as palavras, mas também sobre as combinações de palavras, isto é, sobre o comportamento (as propriedades sintáticas e semânticas) dessas combinações.

Os dicionários de uso, informatizados ou não, não estão sujeitos a essas imposições. Para não sobrecarregar o dicionário, muitas informações evidentes para o utilizador (humano) são omitidas, muitas outras são apenas implicitamente referidas.

Pressupõe-se, em muitos casos acertadamente, que os falantes que os consultam têm conhecimentos lingüísticos suficientes para estabelecer relações e reconstituir o que eventualmente falte. Mas às máquinas é preciso dizer tudo de forma completa, explícita e coerente.

Os dicionários de uso são concebidos para serem usados por humanos, não possuem, por mais completos e bem elaborados que sejam, os requisitos necessários à sua utilização automática.

Os dicionários eletrônicos descrevem as palavras simples e compostas de uma língua, associando a cada uma um lema e uma série de códigos gramaticais, semânticos e flexionais. Esses dicionários, no léxico-gramática, são representados com o formalismo DELA e foram elaborados por equipes de lingüistas para várias línguas (francês, inglês, grego, italiano, espanhol, alemão, tai, coreano, polonês, norueguês, português...).

As várias tentativas para reconverter os dicionários de uso em dicionários eletrônicos, isto é, em léxicos que possam ser usados automaticamente em operações de processamento das línguas naturais, têm-se revelado uma tarefa difícil, uma vez que a explicitação da informação implícita nas definições obriga a reescrever completamente o conteúdo das entradas.

2.5
A importância do léxico de palavras compostas para o processamento

Conforme observou Gross (1988:58),

> *A necessidade dessa pesquisa não se atém apenas ao tamanho do léxico em questão (várias centenas de milhares de elementos), mas à freqüência dos compostos nos textos. Podemos ter uma idéia sublinhando em um jornal ou em obras científicas as seqüências mais ou menos fixas (substantivos ou outras categorias). Nós vamos perceber que é ilusório sonhar com um tratamento automático antes de dispormos de uma parcela considerável de cobertura das estruturas compostas.*

M.Gross e D. Tremblay (1985) afirmam que os substantivos compostos constituem uma boa parte do léxico das línguas. Baptista (1994:2), mostra que as palavras compostas constituem uma parte substancial do léxico de qualquer língua. Os nomes compostos representarão, provavelmente, a maior parte do léxico composto. Nas línguas européias, os substantivos compostos encontram-se na faixa dos milhões. No Brasil, o reconhecimento da extensão do léxico nominal composto é ainda muito deficiente – as palavras compostas que constam nos dicionários usuais ainda não representam sua real existência no léxico.

Silberztein (1997) chama atenção para a importância de ferramentas lexicais, pois, para confrontarmos o dicionário com um texto e associar as palavras do texto às informações lingüísticas do dicionário, precisamos de ferramentas de análise lexical. Nesse sentido, registramos a relevância de uma descrição para as palavras compostas que trate, por exemplo, dessa questão, a fim de que se possam aperfeiçoar alguns dos problemas clássicos no processamento das línguas: armazenamento de dados numerosos, reconhecimento de formas numa seqüência linear sem o comprometimento das idéias no que diz respeito a ambigüidades, redundâncias, repetições, informações incorretas e agramaticais.

Assim a elaboração de um dicionário eletrônico de palavras compostas é uma aplicação factível que poderá auxiliar tanto na resolução de problemas de programas que manipulam a língua, quanto na resolução de problemas de programas que levam em consideração as características dessa língua. Quanto mais recursos houver para a descrição da língua, melhor será a qualidade dos programas de processamento da linguagem natural.

A necessidade de elaboração de um dicionário eletrônico de palavras compostas se configura como um dos recursos que poderão contribuir para a melhoria da qualidade das aplicações. Ou seja, a qualidade de uma aplicação depende muito do dicionário acoplado a ela.

2.6
A delimitação das unidades lexicais compostas

A delimitação das unidades lexicais compostas é um tema que, conforme veremos no capítulo 5º, não é suficientemente resolvido nem pela abordagem das gramáticas tradicionais nem por lingüistas. Embora haja muitos estudos sobre esse assunto, ainda não há um consenso a respeito do conceito de composição e de critérios formais que possam identificar uma seqüência composta de uma seqüência livre, especialmente as que apresentam a estrutura N de N. O estudo das seqüências candidatas a compostos que apresentaremos, abrange dois aspectos: um de natureza lingüística e um do ponto de vista da representação e codificação para fins computacionais. Nesse caso, temos que lidar com algumas dificuldades em torno das múltiplas dimensões do conceito de palavra, que nem sempre coincidem. Não se pretende aqui discutir a noção de palavra, embora seja relevante para compreendermos as diversas definições de composição. Mas para atender aos objetivos desta pesquisa - o de identificação de critérios e codificação das palavras compostas -, destacamos o problema em torno do conceito da não-composicionalidade, pois, para que uma palavra seja considerada composta, deve ser não-composicional. Aí está a grande dificuldade para se decidir entre os tipos de critérios que melhor possam descrevê-las e identificá-las. Do ponto de vista gramatical, percebe-se que há uma preferência por critérios semânticos, e, de fato para aquelas palavras totalmente opacas, do tipo *lua-de-mel* e *jogo de cintura*, se já conhecemos o sentido da palavra, o critério semântico pode esclarecer o sentido de composição. Porém, observamos que nem todas as palavras compostas são do tipo totalmente opacas, pois há casos em que podemos perceber uma certa transparência semântica em um dos componentes da palavra, por exemplo, em *pano de prato*, *fim de semana*, *cartão de crédito*, *toalha de banho,* etc. Há outros seqüências, por sua vez, em que podemos observar ambigüidades, por exemplo, em *rabo de cavalo* (parte do corpo do cavalo) e um (tipo de penteado), *copo-de-leite* (tipo de flor) e *copo de leite* (copo que contém leite). Nesses casos, precisamos recorrer também aos critérios sintáticos e morfológicos para as descrições, sobretudo se se pretende estabelecer

critérios formais.

De modo geral, como veremos no capítulo 5º a noção de composicionalidade está atrelada à transparência semântica dos constituintes da palavra e a noção de não-composicionalidade à ausência dessa transparência. Com base na análise de *corpus,* pretendemos, no capítulo 7º, mostrar que, a partir de critérios que levem em conta as propriedades sintático-semânticas dessas combinações de palavras, uma palavra composta pode ou não apresentar transparência, pois, conforme demonstraremos em nossa análise, cap. 6º esse não é o único critério que pode definir se uma palavra é ou não composta, aliás, esse critério provavelmente só daria conta de explicar aqueles casos de palavras totalmente opacas. Observamos que há um grande número de palavras, com estrutura N de N, que não são totalmente opacas e, entretanto, podem ser compostas.

A identificação de seqüências, com estruturas N de N, como um item lexical, ou seja, como uma seqüência composta, traz para o processamento automático da linguagem natural uma enorme contribuição, pois as informações dadas em configuração lingüística, quase sempre, apresentam dificuldades de interpretação, devido a questões relacionadas à ambigüidade ou à noção de composicionalidade.